UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOSOFÍA

Departamento de Teoría del Conocimiento, Estética e Historia del Conocimiento

TESIS DOCTORAL

Estímulo, significado, consciencia: un estudio sobre los fundamentos de la psicología cognitiva y la eficacia causal de lo mental

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Juan Hermoso Durán

Director

Pedro Chacón Fuertes

Madrid, 2014

ESTÍMULO, SIGNIFICADO, CONSCIENCIA: UN ESTUDIO SOBRE LOS FUNDAMENTOS DE LA PSICOLOGÍA COGNITIVA Y LA EFICACIA CAUSAL DE LO MENTAL

Memoria para optar al título de Doctor por la Universidad Complutense de Madrid, presentada por D. Juan Hermoso Durán bajo la dirección del Dr. D. Pedro Chacón Fuertes, en el seno del Departamento de Teoría del Conocimiento, Estética e Historia del Pensamiento de la Facultad de Filosofía de dicha Universidad.

Madrid, 2014

VNIVERSITAS COMPLVTENSIS MATRITENSIS

ESTÍMULO, SIGNIFICADO, CONSCIENCIA: UN ESTUDIO SOBRE LOS FUNDAMENTOS DE LA PSICOLOGÍA COGNITIVA Y LA EFICACIA CAUSAL DE LO MENTAL

Juan Hermoso Durán

A mis padres

A Pablo y Martín

A Olga

[…] nuestra alma se ufana del privilegio de reducir a su condición todo aquello que concibe, de despojar de cualidades mortales y corporales todo lo que le llega, de obligar a las cosas que estima dignas de su intimidad a desvestirse y despojarse de sus circunstancias corruptibles, y a hacerles dejar de lado, como vestidos superfluos y abyectos, espesor, longitud, profundidad, peso, color, olor, aspereza, lisura, dureza, blandura y todos los accidentes sensibles […], de tal manera que la Roma y el París que tengo en el alma, el París que imagino, lo imagino y lo comprendo sin extensión ni lugar, sin piedra, sin yeso y sin madera.

Michel de Montaigne

ESTÍMULO, SIGNIFICADO, CONSCIENCIA: UN ESTUDIO SOBRE LOS FUNDAMENTOS DE LA PSICOLOGÍA COGNITIVA Y LA EFICACIA CAUSAL DE LO MENTAL

Sumario

ENTRE EL MUNDO Y LA MENTE: LITIGIOS FRONTERIZOS

Captatio beneuolentiae

Las indagaciones que abrigan estas páginas han ocupado, con desigual intensidad, los últimos quince años de mi vida. Estaba ya embarcado en esta investigación cuando me fue por primera vez dada, formalmente, una labor de enseñanza: muchas de las reflexiones que aquí se perfilan provienen del esfuerzo por explicar alguna cuestión difícil a un estudiante que, a diferencia de tantos, tenía la sagacidad de saber que no entendía y el coraje de admitirlo. En el tiempo que ha abarcado esta investigación han nacido mis dos hijos, que me han enseñado tanto más de lo que hubiera podido yo aprender nunca; también la enfermedad y su mirada huera han hecho mella en mi corazón. Las más de las veces, sin embargo, no ha sido la presencia dichosa o terrible de los extremos de la vida lo que me ha apartado de las lecturas o las reflexiones que habían de ir dando forma a estas páginas, sino el ajetreado día a día que conlleva tratar de mantener a punto los engranajes en que descansa el quehacer de una nutrida comunidad de profesores y alumnos. Las demoras que todo ello ha ido imprimiendo a este trabajo, y que tantas veces parecían no tener fin, permitieron quizá, por otro lado, que los pensamientos que en él se plasman bebiesen de multitud de fuentes a las que de otro modo no habría tenido tiempo de acercarme.

A menudo, en estos años, me han preguntado de qué trataba esta investigación. Cuanto más iba adentrándome en ella, más difícil se me hacía contestar, y desde hace ya un largo tiempo vengo eludiendo la respuesta. Creo que va siendo hora: esta investigación trata sobre las diferentes maneras en que podemos intentar explicar las acciones humanas –las acciones de nuestros semejantes y las nuestras propias– y sobre la relación entre esas diferentes maneras; en particular, esta investigación trata sobre si existe una manera de explicar las acciones humanas –presumiblemente, la que articulan la fisiología, la bioquímica y, en último término, la física– sobre la que a la larga hayan de revertir todas las demás o si por el contrario hay otras que puedan reclamar para sí el don de hacer comprensibles aspectos de la acción humana que de otro modo permanecerían opacos; más en particular, esta investigación trata sobre si en las raíces del modo de construir teorías psicológicas que bajo el nombre de *cognitivismo* acabó, hará poco más de medio siglo, con cierta, breve hegemonía del movimiento conductista en seno de la psicología científica, es posible encontrar o no un fundamento sólido para la idea de que la explicación psicológica puede en efecto reclamar tal don para sí, y hacerlo además, como se ha pretendido, sin cuestionar que en última instancia todo cuanto existe en cada uno de nosotros es lo que en última instancia se conforma en su cuerpo; más aún, esta investigación trata sobre si la discutida autonomía de la explicación psicológica depende crucialmente o no de que nuestros cuerpos puedan encontrarse en estados cuyos lazos causales con el entorno que habitamos y con nuestras acciones sólo queden adecuadamente aislados si los describimos en el lenguaje teórico que nos proporciona la psicología. Al intentar desmadejar los muchos hilos que se enredan en esas preguntas, fue cobrando fuerza la convicción de que el concepto de *significado*

formaba la urdimbre sobre la que habían de tejerse las respuestas; después, la de que el concepto de *consciencia* perfilaba el horizonte hacia el que esas respuestas habían de mirar; por último, la de que las raíces del problema, en su despliegue histórico, se intricaban en torno del concepto de *estímulo*. ***Estímulo, significado, consciencia*** sería, así pues, un título acertado para ***un estudio sobre los fundamentos de la psicología cognitiva y la eficacia causal de lo mental.***

Esta investigación nunca tuvo un final, no al menos el final que yo, ingenuamente, había esbozado para ella. Tuvo, a lo sumo, una *capitulación*: me di por vencido cuando la medida del tiempo que le había consagrado parecía ya tan estrafalaria como la extensión de estas páginas. Me temo que sería presuntuoso, en consecuencia, pretender que tuviera una conclusión –a lo sumo podré ofrecer una *re*capitulación. Es ésta: creo que no encontraremos una respuesta convincente a la pregunta en torno al estatus de la explicación psicológica si no logramos antes descifrar el papel que en nuestra noción de causalidad desempeñan las causas mentales y, con ellas, el significado propio de los estados mentales que las conforman y la consciencia de la que de tanto en cuanto vienen revestidos. Pero aclarar por qué he llegado a esta convicción es mucho más laborioso, y llevará sin duda algo más de tiempo.

Agradecimientos

"Παρὰ τοῦ πάππου Οὐήρου", se lee al comienzo del libro I de las *Meditaciones* que a modo de notas para sí mismo –"Τὰ εἰς ἑαυτόν"– dejara escritas Marco Aurelio a lo largo de los años finales de su vida: "De mi abuelo Vero", dice quien sería bautizado por Nicolás Maquiavelo como el último de los emperadores buenos, "el buen carácter y la serenidad". Agradecer es, nos enseña Marco Aurelio, no sólo corresponder al beneficio que se nos ha hecho con la mínima ofrenda de su reconocimiento, sino también obligarnos a no perder de vista la huella que ha dejado en nosotros la generosidad ajena: no sólo algo que decimos a quienes dirigimos nuestro agradecimiento, sino también a nosotros mismos, como recordatorio de los muchos débitos que nos conforman.

Antes de seguir adelante, así pues, debe quedar anotado que sin la paciencia inagotable del profesor Pedro Chacón y sus siempre mesurados consejos –el más importante de los cuales, que acotara férreamente el objeto de la investigación, nunca logré obedecer–, este trabajo, como es obvio, nunca habría llegado a buen puerto; tampoco habría habido, sin su generosidad, primeras publicaciones ni primeras experiencias docentes. Es precisamente su benevolencia lo que hace impensable achacarle ninguna de las tachas que se encontrarán en este trabajo, a la vez que hace obligado reconocerle cualquier virtud que pueda atesorar, pues él sin duda la habrá alentado.

Cuando, sin conocerle, llamé a la puerta de su despacho de la Facultad de Psicología, en la primavera de 1996, para hablarle de la tesis doctoral que entonces tenía pensado escribir –y que en bien poco se parecería a ésta–, venía de pasar no pocas mañanas en un improvisado seminario en lo que era entonces la Sección Departamental de Psicobiología, en el que el profesor José María Velasco nos había adentrado, a mi compañero Adolfo Maldonado y a mí, en el debate contemporáneo sobre filosofía de la mente.

La lectura a la que más tiempo habíamos dedicado en aquel seminario era *El redescubrimiento de la mente*, de John R. Searle, así que cuando en septiembre de 1997, merced a una beca del programa de intercambio académico entre la Universidad Complutense de Madrid y la Universidad de California, llegué al campus de Berkeley, no tardé en matricularme en todas y cada una de las asignaturas que aquel año impartía el profesor Searle. Ya se tratara de un curso introductorio pensado para *freshmen*, ya de un seminario de doctorado, mi idea de la reflexión filosófica fue quedando punteada al escuchar cómo el profesor Searle desplegaba sus argumentos sobre las nociones de intencionalidad y consciencia –razonando en voz alta, volviendo sobre sus pasos para reexaminarlos, esquivando las objeciones de algún alumno, despertando a menudo la sonrisa o incluso la contenida hilaridad que no pocas veces acompañaba a la inconfundible mezcla de rigor lógico y apego al sentido común en que solía descansar su *crítica de la razón cognitiva*. El interés que en él y en su ayudante de docencia, Jennifer Hudin, evocaron los trabajos que en torno al concepto de dolor en los argumentos de Saul A. Kripke hube de presentarles fue

entonces, y volvería a ser en algunos momentos de flaqueza, un acicate para perseverar en esta investigación, pues aquellos no eran sino sus balbuceos primeros. Pero otro tanto podría decir de los profesores Hubert L. Dreyfus y Walter J. Freeman, que exploraban cada semana antes nuestros ojos las lindes entre fenomenología y neurofisiología, transitando con toda naturalidad de la trabajosa descripción del funcionamiento del córtex auditivo a la no menos trabajosa lectura de Merleau-Ponty, y de las penetrantes preguntas con que Sean Kelly iba dando forma a sus reflexiones y a nuestros apuntes; también de las pausadas pero inexorables indagaciones del profesor Barry Stroud acerca de la naturaleza del color –un asunto que me había inquietado desde niño– y de las igualmente inexorables y pausadas matizaciones con que el profesor Bernard Williams iba acechándolo –"*But Barry… don´t you think…?*", o incluso del fugaz paso por Howison Library de Ned J. Block para disertar en torno a la viabilidad de predicar propiedades cromáticas respecto de representaciones mentales, una ya lejana tarde de 1998.

Al otro lado del campus, en Tolman Hall, la reflexión sobre el color servía también de gozne entre lo mental y lo físico, y las restricciones empíricas meticulosamente fijadas por el profesor Stephen Palmer a la posibilidad conceptual de una inversión de las relaciones entre longitudes de onda y experiencias cromáticas –cuestión, como es sabido, muy cara al profesor Block– me ayudaron a entender algo mejor no sólo los tornadizos lazos entre lo concebible y lo posible, sino también el tiento con que es obligado avanzar cuando se conjugan premisas construidas con diferentes vocabularios teóricos. Ese mismo cuidado aprendería a reconocer –ojalá también a remedar– en las investigaciones del profesor John Kihlstrom sobre la naturaleza de la consciencia, en las que la voz de William James o Wilhelm Wundt podía escucharse, distante pero nítida, en una frase entresacada de un artículo de Larry Weiskrantz acerca del síndrome de visión ciega o incluso de un trabajo sobre protocolos de anestesia. Conciencia, experiencias de dolor o de color e intencionalidad conformaban, con todo, una visión un tanto solipsista de la vida psíquica, y sería en el vivo debate suscitado en el seminario sobre teoría de la mente que dirigía la profesora Alison Gopnik donde empezaría a vislumbrar el carácter constituyente de la presencia de los demás en cada uno de nosotros. Después, ya en primavera, las sosegadas reflexiones de la profesora Eleanor Rosch en torno a la construcción de la memoria autobiográfica harían arraigar ese convencimiento, y dejarían sembrados mis apuntes de otras muchas intuiciones que aún espero algún día tener tiempo de explorar.

Aquel año de intensísimo aprendizaje no habría sido posible sin la fascinación que habían sabido suscitar en mí la profesora Susana López Ornat y los profesores Luis Enrique López Bascuas y Fernando Colmenares, pero tampoco sin su apoyo expreso: a sus desconcertantes lecciones sobre el desarrollo del lenguaje y de la visión espacial, sobre la etología del comportamiento social y, sobre todo, sobre las conclusiones que de ello cabía derivar en cuanto hace a la naturaleza de lo mental –desconcertantes, claro, para un estudiante poco acostumbrado a que la labor docente

dejara abiertas tantas preguntas como respondía–, se sumaría luego su generosidad al redactar las tres cartas de presentación que exigía Berkeley.

El regreso a Madrid traería consigo la ocasión de ahondar, de la mano otra vez de Pedro Chacón, en la espinosa cuestión del lugar de la consciencia en los modelos cognitivistas de lo mental, así como de seguir desbrozando el camino que para mí abriera primero Fernando Colmenares y luego Alison Gopnik, ahora bajo la pujante luz de las palabras de Ángel Rivière: su seminario sobre teoría de la mente en la Facultad de Psicología de la Universidad Autónoma de Madrid fue una prolongación del embeleso que había sentido, en Berkeley, al contemplar el ejercicio vivo e insobornable del pensamiento. También nos falta hoy, como Ángel, el profesor Eugenio Fernández, cuya acerada inteligencia de Spinoza y del afán con que el Barroco tratara de domeñar la inquietante ubicuidad de las pasiones –develando, como él certeramente decía, el orden de los afectos– tanto me ayudó a hacerme cargo de que la investigación sobre la naturaleza de la mente, por mucho que se impregne de tintes conceptuales o empíricos, no puede desligarse de la reflexión moral. Aún me dejaría tiempo aquel curso, por último, para afianzar mis desordenadas lecturas sobre representación del conocimiento aprendiendo con la profesora Felisa Verdejo, del Departamento de Lenguajes y Sistemas Informáticos de la Universidad Nacional de Educación a Distancia, a construir un rudimentario sistema experto en el viejo PROLOG.

Unos diez años después, en un seminario sobre sistemas expertos en evaluación psicológica impartido en el Colegio Universitario Cardenal Cisneros, tuve el privilegio de volver a fatigar ese terreno codo con codo con el profesor Luis María Laita de la Rica, de la Universidad Politécnica de Madrid, que con su bondad sin término insistía en hacerme ver la destreza con que lo transitaba. Pero diez años atrás, cuando habían transcurrido ya casi otros diez desde que abandonara los estudios de programación con que mis padres intentaran labrarme un futuro, escribir de nuevo línea tras línea de código se me había hecho tan arduo que no tuve otro remedio que pedir auxilio a un buen amigo suyo de siempre –suyo, digo, de mis padres–, y mío de los días de la infancia, el profesor Rodolfo Fernández. Él antes que nadie, apenas llegado yo a la Facultad de Psicología, había intentado mostrarme el horizonte que de cara a nuestra comprensión de nosotros mismos abrían esas mentes artificiales que se adivinaban en las computadoras. Su abrupta ausencia le ha impedido ver que el fruto de su empeño, aunque escueto, habría de llegar, pero le agradaría, creo, saber que su recuerdo tiñe las páginas de este trabajo.

Durante el largo tiempo que esta investigación se ha demorado, ha sido la confianza inquebrantable de Luis Lázaro la que ha procurado el sustento que, tanto como los seminarios o las lecturas, la ha hecho posible. No menos decisiva ha sido su generosidad al permitirme pasar alguna que otra mañana de trabajo en esta o aquella biblioteca, hojeando tantas referencias pasajeras, o en casa, redactando algún fragmento de estas páginas. Que, con desmedida benevolencia, él viese en mí, a quien había encomendado apenas dos años antes la enseñanza de la hoy extinta *Filosofía de la Psicología* en la División de Psicología del Colegio Universitario

Cardenal Cisneros, a la persona idónea para hacerse cargo también de impartir la hoy maltrecha *Historia de la Psicología* acabaría imprimiendo un giro a esta investigación mucho mayor de lo que yo podía entonces suponer: ya nunca pude desistir de una mirada histórica sobre los asuntos que el desarrollo de los argumentos me abocaba a abordar.

A lo largo de los otoños de 2009 y 2010, además, la oportunidad de impartir en el campus de Madrid de la Universidad de Saint Louis una introducción al pensamiento griego, que debo a la confianza de John R. Welch, no sólo contribuyó a ese sustento, sino que también me permitió afrontar como un gozoso deber lo que hasta entonces había sido una pasión que reservaba para el verano. Inevitablemente, también de aquello pueden encontrarse huellas en este trabajo, desde la lectura del *Crátilo* hasta el modo en que el intento de ahondar en la idea de *lógos* lo suficiente al menos como para poder explicarla me llevó a entrever en su intimidad y su distanciamiento con la nuda realidad las raíces del lugar que la posibilidad del error ocupa, creo, en la noción de intencionalidad.

Por lo demás, he mencionado ya que esta aventura germinó bajo los auspicios de una beca del programa de intercambio académico entre la Universidad Complutense de Madrid y la Universidad de California, pero no que continuó brevemente, durante los primeros meses del curso 1999-2000, bajo los de una beca predoctoral de la Universidad Complutense, a la que renuncié cuando hube de asumir mis primeras responsabilidades docentes.

La mayor parte de estas páginas ha sido escrita en Madrid, pero al releerlas reconozco también largos fragmentos redactados en Molino de la Hoz, en Cádiz, en Rota, o en Madrigal de la Vera, y que son por tanto deudores del cobijo prestado por mis padres –es decir: al margen del modo mucho más hondo en el que todo este trabajo está en deuda con ellos–, o por los padres de mi esposa. Aún sobrevive algún párrafo perdido al que di forma en Berkeley, y algunos más que maduraron en Lima, en la primavera austral del año 2002, merced a la hospitalidad del profesor Ricardo Silva-Santisteban, de la Pontificia Universidad Católica del Perú. Otros fragmentos, que recuerdo anotados a vuelapluma en La Habana, Lisboa, Londres, Venecia o Moscú, hablan más que otra cosa de cómo la investigación que aquí se presenta lleva tanto tiempo engarzada en cada peripecia de lo que ha sido mi vida.

Mi vida, es decir: el amor inconmensurable de Olga Muñoz, que ha dado sentido a todo este esfuerzo –como ha dado sentido a todo lo demás desde que la conocí–; nuestros dos hijos, Pablo y Martín, en quienes ese aliento de sentido cristaliza cada día, tibio e irrepetible; mis padres, de quienes he aprendido todo cuanto en verdad sé –y habría sin duda aprendido mucho más si hubiera sido más espabilado, pues es mucho más lo que tienen que enseñar. Al igual que a cada uno de ellos ha pertenecido cada minuto de este trabajo, les pertenecen sus frutos, exiguos, quizá, y de sabor un tanto extraño. Acaso comenzaba ya a entenderlos Pablo cuando, al ver sobre la mesa de la cocina un tratado de epistemología que yo andaba consultando entonces, dijo complacido: "Papá está leyendo un libro sobre cómo *espistar*".

Exordio

Como la aparición del *Djin* –**El genio de la lámpara**– cuando Aladino acariciara su insospechado tesoro: éste es el símil que Thomas H. Huxley empleó en las primeras ediciones de sus *Lessons in Elementary Physiology* para perfilar la relación entre el surgimiento de un estado de consciencia y la irritación de un determinado tejido nervioso. Un abismo –*chasm*– escribiría poco después John Tyndall que mediaba entre ambos fenómenos; Émil du Bois-Reymond imaginaría un golfo –*Klüft*– alzado "[…] frente a los límites de nuestro ingenio". Estas metáforas de lo inabarcable, de lo incomprensible, forman el espacio del que parte la presente investigación, y se materializan en el resignado *dictum* que el propio du Bois-Reymond pronunciaría, en las lindes del siglo, respecto de la naturaleza y origen de las sensaciones: *Ignorabimus!* La resignación, a su vez, se muestra como uno de los vértices de un campo de fuerzas en el que opera también la frugal modestia de que hacía gala Claude Bernard al excluir de nuestras capacidad de comprensión el porqué de los hechos, así como la tenacidad arrolladora de Santiago Ramón y Cajal, convencido tal vez de que en el ámbito del saber toda rendición es prematura.

La sospecha de que la consciencia pudiera ser un "[…] hecho último de la naturaleza" –éste es el giro que Huxley elegiría, después, para librarse del *Djin*– sigue viva en el debate acerca de lo mental en nuestros días, un debate cuyo tejido parece tensado por las mismas fuerzas y articulado en torno a parecidas metáforas: el hiato explicativo –*gap*– al que alude Levine, o la "[…] llama misteriosa" que parece querer convocar McGinn. Se ha dado, no obstante, un giro de cierta envergadura. Buena parte de nuestros esfuerzos recientes se ha centrado en el intento de entender los lazos entre un pensamiento –o un deseo, o un temor…– y aquello en lo que pensamos –o deseamos, o tememos…–: un trabajo en la estela de la idea de in-existencia intencional en la que Franz Brentano cifró la singularidad de lo mental, salvo en que se acomete dejando entre paréntesis la cuestión de la consciencia. Si du Bois-Reymond creía que el sexto de sus *Welträtsel* –la naturaleza del pensamiento– caería ante nosotros como fruta madura si pudiéramos desvelar el quinto –el surgimiento de la sensación–, nuestro propio empeño, como con ánimo bien distinto hacen ver Zenon W. Pylyshyn, Colin McGinn, Daniel C. Dennett, John R. Searle o Jerry A. Fodor, ha sido perseverar en el asedio de aquél asumiendo nuestra ignorancia respecto a éste.

Actitudes, proposiciones, hechos son, así pues, las madejas con las que se teje la indagación acerca de los lazos entre mente y mundo. Si bien la pregunta por las relaciones entre –digamos– una creencia y aquello que creemos es a todas luces diferente de la que concierne a las relaciones entre la creencia y –por recrear el lenguaje de Huxley– la irritación nerviosa, no es menos obvio que entre una y otra cuestión han de existir pasadizos que valga la pena iluminar: que Brentano consignara la intencionalidad como marca de lo mental puede verse entonces como un modo de advertirnos de la profundidad de dichos pasadizos. Pero es en la idea de Bertrand Russell de que es fructífero pensar en creencias o deseos como *actitudes* que

mantenemos hacia determinadas *proposiciones*, las cuales a su vez se refieren a tales o cuales *hechos*, así como en el marcado giro lingüístico que Roderick Chisholm diera al estudio de estas actitudes proposicionales al centrarlo en el *análisis del comportamiento lógico de los enunciados del lenguaje coloquial que se emplean para atribuir tales actitudes* a otros o a nosotros mismos, donde el cognitivismo ha encontrado la más caudalosa fuente de inspiración para dar forma al abierto *recurso a representaciones internas en la explicación de la conducta* que le sirvió para desligarse de la tradición conductista. A pesar de que la armazón conceptual alzada por Russell o Chisholm queda lejos de proveernos de explicación alguna –pertenece más bien, como se dirá, a la topografía del *explanandum* que a la fábrica del *explanans*; es, si se prefiere, explicativamente inerte–, lo cierto es que al ceñir los cimientos de dicha fábrica deja ya fijados algunos de sus rasgos principales. Así, pongamos por caso, el cognitivismo se aboca a perfilarse como una reivindicación de la psicología que de algún modo se halla implícita en ese lenguaje coloquial –reivindicación cuyos términos habremos de esmerarnos en delimitar. Con ello, asuntos como la proliferación en el seno de dicho lenguaje de contextos intensionales –*i.e.*, refractarios al principio de sustituibilidad *salva veritate* de términos correferenciales que ha quedado consagrado como *ley de Leibniz*– aparecen como claves de las que una teoría psicológica madura debería rendir cuentas. En el desarrollo del cognitivismo resultaría decisiva, en efecto, la idea de que el control efectivo de la conducta compete a las representaciones internas, y no a los estímulos –idea que es, según se verá más adelante, se revela como un trasunto de la de intensionalidad.

Si reemplazar un término por otro que se refiere a lo mismo puede hacer falso un enunciado verdadero –o viceversa–, es razonable pensar que esto ocurra porque el término no se emplee en virtud de aquello a lo que se refiere –su extensión, su denotación–, sino del modo en que lo hace –su intensión, su connotación. Tanto los trabajos lógicos de Aristóteles como los de John Stuart Mill destellan, pues, entre los orígenes de la concepción cognitivista de lo mental, pero mucho más rotunda es sin duda **La sombra de Frege**. Como es bien sabido, en el transcurso de las investigaciones de Gottlob Frege sobre la naturaleza de la relación de identidad, el sentido –*Sinn*– de un signo (o una expresión) se va perfilando como las propiedades semánticas que lo diferencian de otro signo (o expresión) con el (o la) cual comparte una misma referencia –*Bedeutung*–; aquello, por tanto, que permite que un enunciado que una a ambos signos (o expresiones) en torno a un signo de identidad, "=", no resulte forzosamente tautológico. El sentido es, entonces, no sólo aquello que determina su referencia, sino además aquello que aprehendemos cuando entendemos un signo o una expresión. Pero el sentido no puede ser –piensa Frege– una representación interna –una intuición o presentación, *Vorstellung*–: tales representaciones, que atañen a la psicología, pueden variar indefinidamente entre sujetos, pero el sentido de un signo, so pena de hacer imposible toda forma de diálogo, ha de ser estable. Al destilar de esa idea ingenua del sentido como *Vorstellung* todo vestigio psicológico acrisola Frege su noción de *pensamiento*: aquello que asevera una oración afirmativa –un juicio–, y que equivale a su sentido.

Naturalmente, además de expresar un pensamiento –esto es, de albergar un sentido– un juicio bien puede ser verdadero o falso. Comoquiera que ese valor veritativo no es el pensamiento expresado –esto es, el sentido–, Frege lo identifica con la referencia del juicio. Pero sabíamos que es el sentido lo que determina la referencia –y lo que captamos cuando entendemos–; ahora sabemos, por tanto, que el sentido de un juicio porta consigo su valor de verdad. Entre los bastidores de la concepción fregeana del significado, así pues, se opera una exhaustiva purga cuyo propósito no es otro que desproveer a la lógica de cualquier tonalidad psicológica, y cuyas consecuencias para nuestra concepción de lo mental son múltiples y de profundísimo alcance.

Constatamos, por un lado, cómo la verdad o la falsedad de un juicio, que depende de su sentido, han quedado expulsadas de los dominios de la psicología: tal como nos recordaría Kenny, si hubiera leyes que describieran el encadenamiento de estados mentales, éstas no harían "[…] ninguna distinción entre pensamientos verdaderos y […] falsos". De nuevo, es fácil entrever en este punto las fuentes de la primacía que la representación adquiriría en el seno del cognitivismo en detrimento del estímulo –es decir, de los hechos; es decir, de la verdad de la representación. La argumentación de Frege, con todo, ofrece una primera oportunidad de bosquejar una reivindicación de la relevancia, en la explicación psicológica, de los lazos que las representaciones internas traben con el mundo: en pocas palabras, si asumimos que estados psicológicos como las creencias se originan en ese tráfago causal que comienza en la estimulación de los sentidos –aun cuando aceptemos que recibe también el caudal de otros afluentes– *y* no incorporamos una explicación de la posibilidad del error en los fundamentos de nuestra teoría psicológica, ya de poco servirá que intentemos –como el propio Frege– hacerlo después.

Por otro lado, asistimos también en Frege a un riguroso pupilaje de las peculiaridades del lenguaje psicológico coloquial a un caso más general, el de la mera cita: "Duncan creía que Macbeth era digno de confianza" no es entonces esencialmente diferente de "Tales dijo que el agua es el principio de todas las cosas"; en ambos enunciados, lo que la oración subordinada aporta al sentido de la principal –es decir, al pensamiento expresado por el juicio–, y por esa vía a su referencia, no es su propia referencia –es decir, su valor de verdad– sino su sentido, y sólo podría por tanto quedar reemplazada *salva veritate* por otra de sentido idéntico. Así pues, tomar, de la mano de Chisholm, el comportamiento lógico de determinadas expresiones del lenguaje psicológico coloquial como brújula para nuestra comprensión de lo mental aparecería como una maniobra que sólo ha resultado viable al amparo de una lectura de Frege de la que cuidadosa o burdamente se ha segado cualquier retazo de aire antipsicologista –como muestra, por ejemplo, la reinterpretación de la idea del sentido en tanto que modo de *determinación* de la referencia como la de un modo de *presentación* de la referencia, confundiendo así *Sinn* y *Vorstellung*. Ha sido quizá Ullin T. Place quien de forma más certera ha escrutado las limitaciones del giro lingüístico emprendido por Chisholm, su origen –que él cifra en la influencia perniciosa de la distinción entre *saber qué* y *saber cómo* trazada por Gilbert Ryle al hilo de ciertas observaciones pasajeras de Wittgenstein–, y algunos de sus frutos menos apetecidos

–fundamentalmente, la postergación del análisis de estados psicológicos tan cruciales como puedan serlo la creencia o el deseo, pero menos ajustados al rígido esquema de la actitud proposicional.

Entre la convicción de raigambre brentaniana según la cual la intencionalidad distingue a lo mental de lo físico y el infatigable empeño por encontrar una explicación de los fenómenos mentales que podamos incardinar sin fisuras en el edificio de la ciencia natural –encarnado quizá ya en Ramón y Cajal, pero que a efectos del debate contemporáneo cristaliza en Willard V.O. Quine–, se circunscribe **La cuestión del naturalismo**. La idea de *naturalizar la intencionalidad* –de explicarla, digamos, en términos de propiedades que no la presupongan– se perfila hoy a menudo como un tributo mínimo, pero ineluctable, a cierta concepción reduccionista de la ciencia, un tributo expresado en ocasiones mediante el concepto de superveniencia y ligado a la idea de que la relación entre una creencia o un deseo y aquello que creemos o deseamos no puede de ningún modo constituir una propiedad primitiva –elemental, básica– de la realidad, como lo serían –como sólo lo serían– las propiedades que postula la física. Otras veces, la idea de naturalización aparece sencillamente como un canon epistemológico irrevocable, casi como una mera exigencia de transparencia en la explicación. Pero también, claro está, cabe entender el afán de naturalizar la intencionalidad como fruto de una mostrenca obstinación en asemejar cuanto no entendemos a "las cosas", que creemos entender mejor. Así que, como queda claro al hilo de una célebre discusión entre Fodor y Searle, lo que se dirime es a fin de cuentas si a la intencionalidad le cuadra el viejo adagio del obispo Butler según el cual *todo es lo que es y no otra cosa*: si es uno de esos hechos últimos vislumbrados por, si al apelar a ella hemos topado con el lecho rocoso en el que –ya lo anunciaba Wittgenstein– "[…] las explicaciones tienen que terminar […]" o si, por el contrario, apenas hemos nombrado aquello que pretendemos comprender. O, tal vez, si el desencantamiento del mundo que procura el conocimiento científico ha de alcanzar también a todos los reductos de la propia mente que lo ha forjado.

La pregunta se torna, entonces, en la de cuáles podrían ser nuestros **Motivos para quebrar el hechizo**, cuál es la mies que nos aguardan, si es que nos aguarda alguna, si finalmente hubiéramos de rendir el cobijo que habíamos creído hallar en la singularidad de lo mental. Pues bien: quebrar el hechizo empieza a entreverse así como un mal menor, un modo de soslayar una cosecha más aciaga. El mal mayor, claro, no es otro que la perspectiva de que haya que decretar la radical inexistencia de aquello que anhelábamos salvar. Ciertamente, que la intencionalidad –o cualquier otra cualidad que queramos hacer distintiva de lo mental– sea una propiedad última de la realidad o que se derive de otras de naturaleza en algún sentido elemental no son las únicas posibilidades lógicas abiertas: cabría pensar también que sencillamente no exista tal propiedad, ya porque la utilización que de ella hacemos en el discurso filosófico o en nuestras explicaciones ordinarias de la conducta no sea más que una ficción útil, ya que porque, además de ficticia, la noción de intencionalidad se torne perfectamente inservible tan pronto como sepamos construir una más acertada, en el vocabulario de ciencias más básicas. Instrumentalismo y eliminacionismo son, pues,

los polos menos y más severo de una interpretación antirrealista de lo mental que su naturalización, después de todo, nos permitiría al menos rehuir. Las dificultades que afronta el naturalismo pueden leerse entonces –el *locus classicus* de ese giro se encuentra en *Palabra y objeto*, de Quine– como un acicate para promover la abolición sin paliativos del vocabulario psicológico tradicional. Desde esta atalaya, en fin, el paisaje resulta suficientemente lúgubre como para que la naturalización de lo mental se vislumbre como un destello esperanzador, como una forma de humanismo.

La tesis de que no haya otra forma de entender la intencionalidad que incorporarla a un presunto inventario de propiedades últimas de lo real o al de sus derivados –o, al menos, que no haya otra forma de entenderla sin vernos arrastrados al antirrealismo– es puesta en tela de juicio por Terry Horgan, quien considera viable construir una idea de propiedad fundamental que deje indemnes las convicciones naturalistas. La articulación de ese delicado equilibrio requiere, no obstante, conceder a Horgan un conjunto de premisas acerca de los motivos que subyacen a dichas convicciones, la relación entre la intencionalidad y las propiedades elementales sobre las que descansa, la idea misma de propiedad elemental, y la naturaleza de los conceptos humanos en general, en torno a las cuales es fácil sembrar dudas. En particular, el argumento de Horgan depende de la tesis de que cualquier caracterización naturalista de las propiedades elementales en las que en el fondo consiste la intencionalidad nos resultaría *intratable*. Pero el único modo de que eso sostenga sus conclusiones es que la intratabilidad en cuestión no sea asunto de una circunstancial penuria, sino más bien –digamos– de una indigencia constitutiva de nuestro entendimiento, y Horgan está lejos de haber dejado afianzada tal cosa. En realidad, la idea de que el vínculo entre lo mental y lo físico desborda nuestra capacidad de concebir hunde sus raíces en un territorio que nos es conocido –Huxley, Tyndall, du Bois-Reymond. Tan convencido, no obstante, como pudiera mostrarse du Bois-Reymond de que incluso un sabio fáustico habría de rendirse a su *ignorabimus* lo estaría poco después Edward L. Thorndike de lo contrario –Max F. Meyer, en la estela de Thorndike, pronto comenzaría, de hecho, a dar forma, en el seno de un conductismo temprano, a la osada conjetura de la inexistencia de esa realidad mental que otros veían impenetrable: a la equiparación de deseos, anhelos o creencias a fantasmas, dioses o demonios.

Pensar cómo podríamos **Pensar sin pensar** se perfila entonces como el reto crucial al que nos enfrenta la pujanza de esa avidez por abolir lo mental que arraiga en el pensamiento de Meyer. La fuente de la que manan los juicios más severos acerca de la realidad de la mente –o, entre éstos, los más firmemente fundados– se halla en la idea de que el discurso psicológico bien pudiera incorporarse al vocabulario propio de la ciencia en tanto reconociéramos su naturaleza teórica. De ese modo parece que pretendían Rudolf Carnap o Wilfrid S. Sellars –pero también, antes, Carl G. Hempel– proclamar que el positivismo lógico admitía en el seno del saber científico a la psicología, antaño desterrado por Auguste Comte. Ahora bien: si los conceptos que conforman ese discurso psicológico son en efecto conceptos teóricos, no cabe negar entonces que pudieran pertenecer a una teoría tan falsa como

longeva; de ser así, ¿qué otra cosa podríamos razonablemente hacer salvo prescindir de ellos, como ya prescindimos de los espíritus animales o el éter?

Consideraciones de esta índole labraron, sin duda, el humus del que brotaron las dudas de John J.C. Smart –"Jack" Smart– sobre el estatus de realidad de los fenómenos psicológicos, que Ullin T. Place y él mismo habían dado por idénticos a sus correlatos neurológicos. Ante las implacables críticas que la tesis de identidad psicofísica y la noción de análisis temáticamente neutral en la que Smart trataba de sustentarla recibieran de manos de Jack T. Stevenson o Marshell C. Bradley, Smart, en la estela que había trazado Paul K. Feyerabend, no pudo sino escuchar el canto de la sirena y conceder que acaso, después de todo, no nos fuera dado identificar un deseo con un estado del sistema nervioso, sino afirmar la inexistencia de aquel en beneficio de la inequívoca existencia de éste. Aunque el mismo Smart tildaría poco después de veleidades sus titubeos eliminacionistas, otros muchos se han esforzado en tantear las consecuencias que acarrearía la inhabilitación del vocabulario psicológico. Entre los hilos de esa discusión vale la pena detenerse en el que trata de hilvanar Stephen Stich o, poco después, David Braddon-Mitchell y Frank C. Jackson: la naturalización de la intencionalidad, o la inviabilidad de tal empeño, resultan indiferentes –insiste Stich– en lo que atañe al estatus ontológico de ésta, según nos muestran otros conceptos incontestablemente científicos, como el de "fonema" en lingüística o el de "conducta de acicalamiento" en etología, cuya naturalización resulta igual de espinosa; los conceptos psicológicos –aseguran Braddon-Mitchell y Jackson– encuentran su nicho entre las ciencias toda vez que no exijamos que éstas únicamente empleen conceptos referidos a clases naturales, o bien que permitamos que tales clases vengan delimitadas, como vienen los conceptos psicológicos, según criterios funcionales. Asemejar creencias o deseos a fonemas, conductas de acicalamiento o, como hacen Braddon-Mitchell y Jackson, a constelaciones –ya Place había explorado en su día las similitudes entre la naturaleza de los estados mentales y la de los electrometeoros–, en lugar de a fantasmas, dioses o demonios, se perfila así pues como un modo de limar las aristas del eliminacionismo. Otro, quizá más acre, pasaría por mostrar cómo hay más conceptos, tan medulares o casi a nuestra visión del mundo como los de deseo o creencia, que habrían de correr la misma suerte: pocos años después de que Meyer diera el paso de desmentir la realidad de lo mental, Francis G. Crookshank, un médico de Londres, abogaba con vehemencia por la abolición del concepto de enfermedad.

Tal vez con maneras demasiado expeditivas ha tratado Searle de abatir las tesis eliminacionistas haciendo ver que la relevancia o irrelevancia de los conceptos psicológicos en la explicación científica es inocua con respecto a la existencia o inexistencia de los referentes de dichos conceptos, como ocurre con tantos otros conceptos de uso cotidiano para los que no hay cabida en el discurso de la ciencia. El error que subyace al eliminacionismo residiría entonces, como ha señalado John Heil, en la identificación de los conceptos psicológicos como parte del *explanans* de una teoría –movimiento que proviene de Carnap, y cuya impugnación Lycan ha ligado al pensamiento de Sellars –, y no como parte del *explanandum* que una ciencia madura

ha de abordar. Hay, desde luego, otros ensayos de aplacar los conatos eliminacionistas, como el de convertirlos en gestos auto-refutatorios que, al defender la inexistencia de creencias, implicarían la imposibilidad de creer en su propia verdad –según intenta Heil–, o el de hacer de ellos palabrería estéril, trivialmente verdadera o trivialmente falsa según cuál sea la teoría de la referencia que adoptemos –con el que Stich de desliga de sus anteriores requiebros con la abolición del vocabulario mentalista. Pese a la encendida controversia que a menudo se ha desencadenado en torno a estas cuestiones, no es difícil reparar en que las tesis eliminacionistas han ido colonizando cierta oratoria sobre lo mental y lo cerebral, aún a costa de convertirse más de una vez en aseveraciones tan solemnes como nimias, en las que la contradicción aflora casi a simple vista; otras veces, en cambio, en la tensión que provoca la presencia de lo inexplicado cobra vigor el mismo aliento poético que desde los tiempos de du Bois-Reymond, y antes, ha venido impulsando no pocos avances científicos.

Como en las fábulas de antaño, la **Quimera de la nube equivocada** nos enseña el modo en que el intento de entender que nuestras palabras o nuestros pensamientos puedan designar o describir *erróneamente* el mundo ha ido entrelazándose con el propio intento de entender que nuestras palabras o nuestros pensamientos puedan, sin más, designar o describir el mundo. De hecho, como veremos, la posibilidad del error se ha ido erigiendo recientemente como la clave que habría de permitirnos dar cuenta de la relación entre la mente y el mundo, en tanto que nota que diferenciaría lo propiamente semántico del signo natural. A modo de coda de estas secciones de aire propedéutico, se hace, entonces, irrefrenable la tentación de articular, aun muy deslavazadamente, un relato de cómo esa distancia que entre las cosas del mundo y los pensamientos o palabras con que tratamos de apresarlas entraña el error ha ido abriéndose paso en nuestra comprensión de nosotros mismos. Hay en ese relato una transparencia originaria, que desde la metafísica biblíca en virtud de la cual el Apocalipsis puede aludir a la muerte de un nombre –por la de quien es por él nombrado– alcanza hasta las conversaciones entre Agustín de Hipona y su querido Adeodato, y que apresta también el trasfondo sobre el que se va dibujando en el pensamiento griego la paulatina consciencia de que, si bien *lógos* es tanto el orden fundamental oculto en el mundo como el discurso o la razón que lo develan, entre esas dos orillas suyas media a menudo un ancho cauce. Así, en el desdén que Heráclito comparte con Parménides hacia "las opiniones de los mortales" encontraremos los primeros destellos de la minuciosa indagación sobre el error que se despliega en el diálogo platónico entre Sócrates, Hermógenes y Crátilo al que éste último da nombre. Ese hiato entre pensamientos y cosas, entre palabras y cosas, habrá de abocar a Platón a una acerba renuncia al lenguaje como norte de los pasos del *philó sophos* –"el más profundo dolor", según expresión de Giorgio Colli, se escondía en la constatación de la pobreza del lenguaje, y las sospechas que ello arrojaba sobre la propia razón. Ese hiato habría de conducir también, a la larga, a la perplejidad moderna ante aquello que comenzó mostrándosenos claro y diáfano: los lazos entre el pensamiento, las palabras que lo expresan y las cosas que designa, que,

mucho después, acabarían por verse –así, tempranamente, en Thomas H. Pear– casi como la esfinge que guarda todos los secretos de la psicología. Muy medularmente, entonces, dicha perplejidad es también la perplejidad, que con inigualable lucidez expresara Wittgenstein, ante el modo en que la posibilidad misma del pensamiento o el lenguaje –por no decir del conocimiento– parecen descansar sobre la posibilidad del error.

No es posible entender siquiera vagamente la reflexión contemporánea acerca de la naturaleza de lo mental y de la explicación psicológica sin hacerse cargo de lo que ha supuesto en este ámbito el movimiento conductista. El *tópos* de la **Crisis y vigencia del conductismo** perfila una breve hegemonía –entendida a menudo como enfermedad de juventud de la psicología– a la que habría seguido un súbito desplome tras el que nada, salvo ciertos hábitos de higiene metodológica, habría quedado en pie. No es difícil, sin embargo, encontrar reconstrucciones más juiciosas del proceso, en las que figuran también la posterior reparación de algunos de los planteamientos de los conductistas –en un esfuerzo por desgranar lo más clarividente entre cuanto pudiera haber en ellos de obcecado–, o, como se verá más adelante, profundas y vigorosas vetas de continuidad entre dichos planteamientos y la concepción cognitivista de la mente, que, según el relato canónico, habría venido a reemplazarlos. Es preciso, además, tener presentes las fluctuaciones en los presupuestos epistemológicos que acerca de las peculiaridades de la explicación psicológica y su relación con otras modalidades de explicación científica agitaban el subsuelo de la comprensión de lo mental, tanto en el seno del propio conductismo como en la transición hacia el cognitivismo. Así, será obligado atender a la relación entre el pensamiento de John B. Watson y el positivismo lógico –que dista mucho de ser la de buena vecindad, pues Watson se aferra a una epistemología de aire comtiano que resulta ya obsoleta para el propio Hempel, y éste se cuida mucho de ligar la suerte del positivismo lógico a la del programa experimental de Watson–, a la renuencia de Burrhus F. Skinner –para quien no hay más lógica de la ciencia que la ciencia de la conducta de los científicos– a aceptar toda epistemología que no sea un escueto inductivismo no ya comtiano, sino baconiano, o al papel de Meyer y del físico Percy W. Bridgman como arquitectos de los puentes entre conductismo y positivismo lógico que luego transitarían neoconductistas como Edward C. Tolman y Clark L. Hull. A todo ello debe añadirse, desde luego, el recuento de las numerosas anomalías que el conductismo iba viendo germinar en su propio seno –las más estrepitosas, tal vez, las que acabarían enfrentando a Karl S. Lashley con Watson a cuenta del problema del control central de la conducta, y a Keller y Marian Breland con Skinner a cuenta de la utilidad de los principios conductistas fuera del laboratorio–, así como el de las diversas presiones externas que cuestionaban su credibilidad –como el desarrollo de la teoría de la disonancia cognitiva por parte de Leon Festinger, la influencia de Kurt Z. Lewin en el seno de la psicología social, o el vertiginoso desarrollo teórico y tecnológico que, de la mano de Herbert A. Simon, Alan Newell o John McCarthy, había de propiciar el concepto de procesamiento de información. Al hilo de todas estas consideraciones, es de rigor, además, hacer

hincapié en la inmensa heterogeneidad de los planteamientos de los propios conductistas –la "torre de Babel conductista" que acertadamente describe Leahey–, que, emborronada por la historiografía cognitivista, se hace imprescindible perfilar mínimamente de cara a una cabal comprensión de buena parte de los problemas que acotan la reflexión actual sobre lo mental. Bajo el prisma, por último, de una revisión de la temprana y duradera polémica acerca de si la transición del conductismo al cognitivismo en psicología constituye una revolución científica en el sentido acuñado por Thomas S. Kuhn, se hace preciso abordar también cuestiones como la continuidad de los planteamientos mentalistas en la psicología europea durante los años de auge del conductismo, el papel de los intereses bélicos o de otras fuentes de apoyo institucional en el ímpetu del cognitivismo, o la propia regularidad y elegancia de ciertos resultados experimentales cosechados en los laboratorios conductistas como acicate de la teorización cognitivista.

En definitiva, frente al lugar común que dicta a un tiempo, sin aparentemente advertir contradicción alguna, que el conductismo sucumbió víctima de su propia, descomedida severidad metodológica y que es en los principios metodológicos donde se observa más claramente su pervivencia en las entrañas del cognitivismo, todo esto nos abocará a la conclusión de que la crisis del conductismo no atañó tanto a sus directrices metodológicas como a sus supuestos teóricos o, quizá más exactamente, preteóricos –aunque, como ya dejara apuntado Yela, a esa crisis de supuestos teóricos subyaga el cuestionamiento de ciertos principios metodológicos, primero, de los principios de interpretación de los resultados experimentales, después, y, sólo entonces, de la naturaleza del objeto de estudio. La piedra angular sobre la que había de construirse la nueva psicología cognitiva –que el propio Skinner reconoció con notable perspicacia–, su núcleo preteórico, no es otra que la idea de que lo que controla la conducta de los organismos no es el entorno sino la representación que se forman de ese entorno: la idea, pues, de que el organismo habita un entorno intencional –o un mundo nocional, si queremos reemplazar el vocabulario de Charles Taylor con el de Dennett. Pero esa idea nos remite de nuevo irremediablemente al terreno ya hollado de la necesidad de rendir cuentas de la posibilidad de que alberguemos representaciones erróneas del mundo –es decir, de explicar la normatividad de los estados intencionales–, y anuncia, además, el ancho horizonte que abre la pregunta por el papel que tales estados intencionales puedan tener reservado en la determinación de las causas del comportamiento.

En el empeño por **entender qué aprendemos** cuando aprendemos –a reconocer ciertas formas, a tararear una melodía, a hablar…– el cognitivista habría luchado entonces por denunciar la **Fingida austeridad** del conductismo, mostrando la penuria explicativa que ocultaba. Los argumentos que con mayor vigor impulsaron la teorización sobre representaciones internas –los de Noam Chomsky y Jerry Fodor– compartían la idea de que el entorno del organismo no basta por sí solo para dar cuenta ni de nuestra capacidad de aprender un lenguaje –como Chomsky reprochaba a Skinner– ni de nuestra capacidad de aprender otras destrezas en apariencia mucho más sencillas –como Fodor desgranaría en su disputa con Ryle,

como Lashley había hecho ya, en diferentes términos, en su litigio con Watson–, y enlazaban sin ambages esa necesidad de cartografiar el territorio que separa al estímulo de la representación interna con la *terra incognita* que Miller, Galanter y Pribram, en la estela de Edwin R. Guthrie, habían sabido adivinar entre la representación interna –el mapa cognitivo– y la conducta. Ver que aquello que otros dan por entendido clama en realidad por una explicación se perfilaría, así, como el signo último del giro que el pensamiento cognitivista imprimiría a la psicología científica. Ante esos gestos de tesón veremos alzarse las ya casi inertes advertencias de Malcolm, de claras raíces wittgensteinianas, de que mudar al reino de lo mental nuestras herramientas explicativas conlleva un grave riesgo de artificio y mistificación, que podríamos esquivar si no desoyéramos la enorme riqueza de tales herramientas que, pese a que las ignorasen Chomsky o Fodor, nos ofrece el entorno en el que se desenvuelve el organismo. Nada podría, en efecto, la exhortación de Malcolm a volver a mirar fuera después de las devastadoras críticas de Chomsky al uso vacuo y subrepticio de nociones mentalistas en el análisis del aprendizaje lingüístico, presuntamente ceñido al vocabulario de estímulos y respuestas, que había forjado Skinner: la mera homonimia entre los términos definidos en el trabajo experimental y los que obraban en dicho análisis, el empleo ritual de la jerga del laboratorio para usurpar la fisonomía de una teoría científica madura, la incapacidad para abordar la cuestión de la intencionalidad siquiera en los casos más sencillos –la utilización de un nombre propio para designar algo que se encuentra ausente del campo estimular–, o, en suma, el dilema entre la irremediable ambigüedad que viciaba las formulaciones skinnerianas bajo una interpretación amplia de su terminología teórica y la lastimera irrelevancia que inexorablemente las infectaba si se hacía de ellas una interpretación más estricta… todo hacía indefendible la resistencia a postular procesos y estructuras internas. Rastrear las huellas que dejarían en la teorización cognitivista los planteamientos de Chomsky –desde Miller, Galanter y Pribram hasta Pylyshyn– es probablemente una tarea inabarcable, pero podremos al menos aprestarnos a ella, algo mejor guarnecidos, indagando primero en las raíces de dichos planteamientos: la polémica sobre la validez de las máquinas markovianas como modelos de la producción lingüística humana, el pensamiento de Lashley –de quien Chomsky se reconocía abiertamente deudor–, pero también ciertas propuestas de Verplanck o de Scriven, y, desde luego, las objeciones de Geach y Chisholm a los análisis disposicionales de creencias y deseos adelantados por Ryle, objeciones en las que cobraría forma la noción de círculo de lo mental que habría de acabar con el conductismo lógico.

Ahora bien: las dificultades que el conductismo afrontaba en su pretensión de articular una explicación global del comportamiento humano exenta de toda alusión a lo mental se manifestaban, casi con tanta claridad como por boca de sus críticos más destacados, en las múltiples **Divergencias y oscilaciones** que se producían en su seno, y que conforman de hecho **las fuentes freáticas del funcionalismo**. Que ya en el manifiesto de 1913 Watson se refiriese al conductismo como una variedad de funcionalismo, aludiendo a cuanto en su llamamiento a una nueva psicología

provenía de las enseñanzas de James R. Angell, señala un sendero a lo largo del cual hemos de encontrarnos también con Watson, Skinner, Weiss, Meyer, Tolman o Guthrie. Así, distinguiremos matices en los que cabe presentir el desarrollo del cognitivismo en la decidida defensa de la autonomía de la psicología frente a la fisiología que Watson empuñaría ante Jacques Loeb, y en la que, por influencia de William J. Crozier, habría de embarcarse también Skinner, al igual que en la insistencia de Skinner en proporcionar definiciones netamente funcionales de estímulos y respuestas –que si pueden figurar en la explicación de la conducta, diría Skinner, es en tanto que *clases* de estímulos y respuestas, definidas según cierto "nivel de restricción"–, pese a su obstinado rechazo a aplicar abiertamente tales definiciones a estados internos y su consiguiente proclividad, denunciada por Chomsky, a hacerlo furtivamente. Aunque ese rechazo no era compartido por Max F. Meyer, que abogaba por la introducción de conceptos psicológicos como abreviaturas de procesos nerviosos complejos –haciendo así patente que sus presupuestos epistemológicos estaban más cerca de los que venía auspiciando el Círculo de Viena que los de Watson o Skinner, anclados en Comte cuando no en Bacon–, la determinación con que tanto él como Albert P. Weiss buscarían el modo de conciliar el conductismo con un matizado reduccionismo de lo mental a lo fisiológico evoca también vívidamente las preocupaciones de los primeros cognitivistas. De la misma manera, que Mayer cifrara su énfasis en la dimensión social de la conducta en la tesis de que lo biofísico y lo biosocial constituyen criterios diferentes de *clasificación* de los procesos sensoriomotores –y, más aun, que supiera ver en ello un modo de articular diferentes vocabularios teóricos sin dejarse arrastrar por el dualismo–, hace de su pensamiento un precedente tan rotundo de las ideas capitales del funcionalismo como lo pueda ser, bien a su pesar, el del propio Skinner. La naturalidad con que Tolman o Guthrie arrostraran la utilización de conceptos mentalistas bajo la forma de constructos teóricos es solamente el más tardío, y quizá también el más conocido, de estos afluentes que el cognitivismo recibe de la concepción conductista de las explicaciones psicológicas. Junto a planteamientos irremisiblemente lejanos de los que darían forma al cognitivismo, cabe, en definitiva, encontrar también en el conductismo, incluso en sus variedades más hostiles a la teorización sobre procesos o estructuras internas, intuiciones en las que dicha teorización queda prefigurada con llamativa nitidez. No sólo, eso sí, se roturaban ya los surcos que habría de transitar el cognitivismo en la agudeza de algunas intuiciones conductistas, sino también en la torpeza de otras: de la notoria ambigüedad, por ejemplo, con que Watson o Skinner tratarían de acotar la lectura ontológica de sus tesis, zigzagueando una y otra vez entre posturas reduccionistas y eliminacionistas cuando no, inadvertidamente, refugiándose en un peculiar compromiso con el epifenomenismo, se alimentaría sin duda la exigencia de esclarecer las relaciones entre nuestra idea de lo mental y nuestra idea de la explicación psicológica que sería característica del incipiente cognitivismo. Sea como sea, parece claro que la concepción de la mente y de su estudio científico que habría de reemplazar al osado proyecto que Watson presentara

en 1913 se encontraba ya en gran medida forjada en el propio seno de dicho proyecto, tal como éste se fue desarrollando en las décadas posteriores.

Las constantes oscilaciones de Watson o Skinner en cuanto a los compromisos ontológicos que entrañaba su concepción de la psicología contribuyeron a hacer del pensamiento de Ryle, notablemente más firme a ese respecto, un eje primordial en el descrédito del conductismo y el avance del cognitivismo. Las dificultades que atenazaban al ensayo de traducción de cualquier enunciado sobre estados o procesos mentales a un conjunto de enunciados sobre conductas o disposiciones a la conducta, tal como Ryle lo había hilvanado, formarían buena parte de la urdimbre sobre la que se tejería el cognitivismo. En particular, dos eran los núcleos problemáticos: la incalculable cantidad de acotaciones referidas precisamente a estados mentales que cada presunta traducción conductual parecía ocultar en su seno, y la ineludible pregunta por el fundamento categórico de las disposiciones a la conducta que figuraban en dichas traducciones. **Fragilidad, dolor**, solubilidad, o la simple creencia de que va a llover se convirtieron en paradigmas contrapuestos de un análisis que se iría antojando cada vez más impracticable: el que se libraba entre **el conductismo lógico y la naturaleza de las disposiciones**. En el trasfondo del debate cobraría un enorme relieve la cuestión de si un determinado estado mental puede darse en ausencia de las conductas o incluso de las alteraciones fisiológicas que habitualmente lo acompañan, una cuestión que contribuiría a precipitar el declive del conductismo a través de un célebre *Gedankenexperiment* sobre el dolor propuesto en 1963 por Hilary Putnam –aunque anticipado por Hempel casi tres décadas atrás– , pero que venía ocupando ya la reflexión psicológica desde que William James expusiera en 1884 su atrevida hipótesis sobre la relación entre las emociones y lo que común –y, a juicio de James, erróneamente– llamamos su expresión corporal. Si las intuiciones de Putnam, *pace* James, eran correctas, tendríamos ubicada la tara que vicia los cimientos del conductismo lógico: la confusión entre los efectos de un estado mental –sus manifestaciones, sus signos…– y sus constituyentes –o, si se prefiere, entre relaciones causales y relaciones lógicas. Pero incluso si fuésemos capaces de delimitar una determinada disposición conductual que pudiera vincularse sin fisuras a un determinado estado mental (y de hacerlo sin mencionar otros estados mentales), seguiría siendo más sensato –piensa Putnam– identificar el estado mental con el estado del organismo que explica tal disposición que con la disposición misma. La idea de que nuestra vida mental no sea sino una sucesión de disposiciones conductuales sin sustrato categórico, que ya había sido rechazada por Geach, conduciría de la mano de David Armstrong a la madurez de la tesis de que los estados mentales son más bien los estados fisiológicos que sustentan tales disposiciones, y en esa confluencia de conductismo y teoría de la identidad psicofísica germinaría el funcionalismo.

La controversia, sin embargo, no cesó en ese punto: Place, por una parte, ensayaría tiempo después una reivindicación de la postura de Ryle que pasa por analizar el papel epistemológico de la noción de disposición distinguiendo entre formas válidas y formas tautológicas de la explicación por *virtus dormitiva*; el propio

Putnam, además, había dejado abierta otra veta de debate al argumentar que la explicación psicológica de la conducta es autónoma respecto de su explicación neurofisiológica en el mismo sentido en que la explicación geométrica de las propiedades mecánicas de un sólido lo es respecto de una explicación en términos de física de partículas –analogía que, como supo ver Elliott Sober, se presta a una interpretación reduccionista contraria al ánimo de Putnam, o incluso a una conductista, que Ned Block trataría de limar. Conviene, con todo, adelantar que tanto en las conclusiones de Place como en las de Sober encontraremos motivos razonables para matizar algunos aspectos de la concepción funcionalista de lo mental que subyace al cognitivismo, como su compromiso anti-reduccionista, o para rehabilitar ciertas facetas del conductismo lógico en las que dicha concepción se hallaba ya prefigurada, pero no para una impugnación *in toto* de aquélla ni para una redención de ésta. Incluso Place, en efecto, admite que el análisis de Ryle partía de una comprensión confusa de las relaciones entre la forma condicional de un enunciado y la atribución de relaciones causales que pueda implicar, y su defensa de Ryle frente a los argumentos de Martin, si es que permite a Ryle esquivar el problema que suponía la pregunta por el fundamento categórico de las disposiciones a la conducta, lo aboca al mismo tiempo al otro atolladero en el que se vio atrapado el conductismo lógico: el ingobernable comportamiento de unos estados mentales que reaparecían aquí y allá, imprevisiblemente, tan pronto como se intentaba proporcionar una traducción conductual de uno de ellos.

La terquedad con la que reaparece el vocabulario mentalista en los análisis conductuales es lo que solemos conocer como el problema del retorno de lo mental. Comoquiera que el funcionalismo puede verse en gran medida como un intento de hacerle frente, y que voces tan vigorosas como la del propio Putnam han alertado de que dicho intento podría no haber sido del todo logrado, quizá sea prudente hablar de, al menos con carácter tentativo, **El (retorno del) problema del retorno de lo mental**. Lo que en 1957 hicieron ver Chisholm y Geach es que incluso la traducción al vocabulario conductual de un enunciado psicológico relativamente sencillo –en el ejemplo de Ryle elegido por Geach como blanco de su crítica, "El jardinero espera que llueva"– sólo es viable en la medida en que una cantidad indefinida de condiciones relativas a otros estados mentales –como que el jardinero no *desea* arruinar el jardín– se asumen de forma tácita o se introducen subrepticiamente en la traducción. Salvo tal vez –apuntaría Chisholm– en el caso de enunciados acerca de la intención de llevar a cabo acciones corporales básicas, como abrir los ojos, no habría modo entonces de dilucidar el contenido de esas cláusulas *caeteris paribus* sin cuya compañía el análisis ryleano resultaría sencillamente falso –y en cuya compañía, por tanto, irremisiblemente vago. El conductismo lógico, en suma, estaba condenado a la circularidad –más aún si, como argumentaba Putnam, no era ya la mención de *otros* estados mentales lo que viciaba el análisis conductual de un estado mental cualquiera, sino, a la larga, la del propio estado mental analizado.

En las objeciones de Chisholm a Ryle ha sabido ver Georges Rey una crítica que cabe extraer del ámbito del conductismo lógico y trasladar a los conceptos clave

del conductismo psicológico, incluso en sus variedades más abiertas a la teorización sobre estados y procesos mentales, como la auspiciada por Tolman. Es razonable argumentar, sin embargo, que ya en la reseña de *Conducta verbal* con la que Chomsky –mucho antes de que cristalizara la propuesta de Rey– había desbaratado la ambición skinneriana de subsumir la explicación toda de la conducta en sus descubrimientos sobre el condicionamiento, la huella de Chisholm y Geach era más que pronunciada, o, al menos, que el problema del retorno de lo mental puede entenderse como la formulación más general y más temprana de los argumentos de Chomsky contra Skinner. Así, por ejemplo, se desprende con claridad del escrutinio de los argumentos que Zenon W. Pylyshyn presentaría en su influyente defensa de la teorización cognitiva frente a las restricciones estipuladas por el conductismo, dirigida contra Skinner pero construida sobre un armazón prácticamente idéntico al de los razonamientos de Geach y Chisholm. En el problema del retorno de lo mental reposaría, vista la cuestión con estos ojos, la lección fundamental que el cognitivismo, de acuerdo con Fodor, habría de aprender de la ruina del conductismo: el carácter relacional de lo mental.

Entre develar la circularidad oculta en la concepción conductista de lo mental y construir una concepción de lo mental purgada de esa circularidad hay un trecho, claro está, que no se recorre sólo con hacer explícito lo que era implícito. Los propios conductistas –Ryle o Skinner sin ir más lejos– habían vertido además duras acusaciones de circularidad contra las aproximaciones mentalistas a la psicología. Pero si del fracaso del conductismo habían aprendido los psicólogos cognitivos que los estados y procesos mentales son esencialmente de índole relacional, del incipiente desarrollo de la teoría de autómatas –de la lectura de los trabajos de Alan M. Turing, en definitiva– habían de aprender, entre otras cosas, la poderosa herramienta que proporciona la idea de *definición simultánea* de cada estado computacional de un sistema en virtud de sus relaciones con todos los demás. Si los estados mentales se identificaban como estados computacionales, el formalismo de la definición simultánea podría mantener a la psicología cognitiva a salvo del círculo de referencias a lo mental que había plagado los análisis conductistas, aunque fuese mediante el expediente, aparentemente precario, de incorporarlo íntegro a sus esquemas explicativos. Años después, un severísimo juez de sí mismo como es Putnam dictaminaría que en esa promesa de la definición simultánea –cuyo cumplimiento, como el de los viejos análisis conductistas, siempre acababa postergándose– se encerraba uno de los males congénitos que a su juicio acabarían con el cognitivismo: una arrogante y desmedida ambición explicativa en la que acaso quepa ver también parte de la herencia conductista de aquella nueva ciencia de la mente.

Aunque el conductismo era en buena medida heredero del positivismo lógico y el operacionalismo –cuando no del pensamiento positivista anterior al Círculo de Viena–, el tenaz rechazo que mostraba, al menos en su vertiente ryleana, a conceder a los estados mentales un fundamento categórico sobre el que hacer descansar la naturaleza disposicional que le era atribuida lo hacía revelarse como un hijo díscolo.

El regreso a los predios del más severo fisicalismo habría de comenzar de la mano de Ullin T. Place, quien identificaría ciertos aspectos de nuestros estados mentales –su componente nudamente experiencial: las llamadas "sensaciones crudas", cuya existencia episódica, *hic et nunc*, se compadecía mal con el análisis en términos de disposiciones–, como estados neurofisiológicos, y dotarlos así de un intachable expediente en términos de eficacia causal. Después, David Armstrong haría por ampliar el radio de la identificación entre lo mental y lo cerebral hasta abarcar también el terreno en el que se había gestado la interpretación ryleana: el de las creencias y los deseos. En este reverdecer del fisicalismo que conlleva la tesis de identidad psicofísica ha querido verse en ocasiones una corriente que confluiría con la del entonces incipiente funcionalismo, pero es más acertado buscar las fuentes de la concepción funcionalista de la mente en las restricciones que paulatinamente se fueron oponiendo a la generalidad de los planteamientos de Place o Armstrong, que en esos propios planteamientos. Se trata, pues, de calibrar el **Despliegue y alcances del fisicalismo**.

En efecto, ya en los trabajos pioneros de Smart o Armstrong se atisban aquí y allá leves, remisos matices a la idea de que defender que la mente no es otra cosa que el cerebro exija hallar para cada uno de los estados mentales que pudiéramos albergar un estado cerebral tal que todo aquél que se encuentre en el estado mental en cuestión, y nadie más, se encuentre en el estado cerebral en cuestión. En los acerados análisis de David K. Lewis, esos matices precipitaron como la distinción entre el ocupante de un determinado rol causal –que bien puede ser un estado cerebral– y el propio rol –con el que cabría identificar el estado mental aparejado. Pero en las enmiendas de Putnam al fisicalismo, tan firmes como lo habían sido sus objeciones al conductismo, se convirtieron en una relectura radical: tal como venía siendo formulada, la tesis de identidad psicofísica se desplomaría con tan sólo el hallazgo de un sujeto –ya fuera un organismo de cualquier especie, una máquina o un desacostumbrado ser angelical– que se encontrara en un determinado estado mental y no en el estado físico que la teoría dictase. Sin embargo, la severidad de la tesis era, a juicio de Putnam, superflua: el mismo compromiso naturalista que pueda derivarse de la afirmación de que albergar un estado mental de un tipo determinado entraña albergar un estado cerebral de un tipo determinado se cosecha también de la afirmación –más moderada, no tan inerme– de que albergar un estado mental determinado –de cierto tipo, claro– entraña albergar un estado cerebral –también claro, de cierto tipo, pero no necesariamente del mismo para todos los estados que resultaran ser del mismo tipo desde el punto de vista mental. Mi dolor y tu dolor, entonces, son dolor en tanto que pertenecen al mismo tipo de estado mental, definido mediante criterios psicológicos –esto es, funcionales–; ambos son también estados físicos –neurofisiológicos, según parece–, pero pueden pertenecer o no pertenecer al mismo tipo de estados físicos, definidos mediante criterios físicos. Era, en suma, una sencilla acotación del alcance de la tesis lo que se reclamaba: abandonar la afirmación de que *para todo estado mental existe un estado físico tal que para todo sujeto*, si el sujeto alberga dicho estado mental alberga también dicho estado físico, y viceversa, y

reemplazarla por la de que *para todo estado mental y para todo sujeto existe un estado físico tal que* si el sujeto alberga dicho estado mental alberga también dicho estado físico, y viceversa. Pero el paso de una tesis de identidad psicofísica formulada entre tipos de estados (o propiedades) –es decir, con alcance general o de tipos– y una tesis de identidad psicofísica formulada entre casos de estados (o propiedades) –con alcance particular o de casos– franqueaba así el camino hacia una concepción de lo mental capaz de simultanear la idea de que los estados mentales exigen su propio nivel de descripción y explicación con la de que no son en último término otra cosa que estados físicos de los seres que los abrigan. Un naturalismo sin reduccionismo, que –insistiría Fodor– es un naturalismo más robusto: el funcionalismo.

Lo que se adivinaba en el horizonte de la reflexión sobre la naturaleza de la mente era, al fin y al cabo, un modo de **Nadar y guardar la ropa: conductismo, fisicalismo y teoría de autómatas** podían engranarse para preservar a un tiempo la naturaleza inherentemente relacional de los estados mentales que habíamos aprendido del conductismo, la impoluta eficacia causal de la que al dotarlos de un sustrato categórico los guarnecía el fisicalismo, y la ductilidad que les daba su conceptualización bajo el prisma de los autómatas abstractos, en la que parecía prosperar el anhelo de un nivel de explicación propiamente psicológico, soberano respecto de la descripción de mecanismos fisiológicos. Dicho engranaje comienza a articularse en la lectura de Putnam de las implicaciones que guardaba de cara a nuestra comprensión de lo mental el trabajo de Turing –en particular, su caracterización de las *máquinas lógicas* como autómatas abstractos cuya naturaleza viene definida por la *tabla de máquina* que especifica su función de transición, más allá del modo en que en cada caso vengan materializados los dispositivos de entrada, memoria y salida de la máquina, o la propia tabla. Entender, pues, qué es exactamente una máquina de Turing se revelará como un trance ineludible para hacerse cargo de la concepción funcionalista de la mente que subyace a la psicología cognitiva.

Que los estados mentales de un organismo pudieran equipararse, en una primera aproximación, a los estados de tabla de máquina de un autómata abstracto, a la vez que abría un nuevo modo de entender los numerosos ensayos de simulación mecánica de comportamientos aparentemente mediados por procesos cognitivos que venían floreciendo desde algún tiempo atrás, dejaba en el aire la pregunta de si el viejo desiderátum conductista de purgar el vocabulario de la psicología científica de referencias mentalistas se había visto por fin consumado. El intento de dirimir la controversia sobre si la concepción funcionalista de la mente entraña un compromiso con la existencia de estados y procesos propiamente mentales, y con el papel de estos en la explicación de la conducta, o si por el contrario constituye más bien un ensanchamiento del proyecto conductista de prescindir de todo ello, articulado ahora en el lenguaje lógico-matemático de la teoría de autómatas, nos exigirá una fugaz profundización en el procedimiento de definición de términos teóricos ideado en Cambridge por Frank P. Ramsey –que, desplegado luego de la mano de Rudolf Carnap y David Lewis, ha cobrado carácter canónico en el seno del funcionalismo–,

así como en las nociones de sistema primario y sistema secundario de una teoría científica articuladas por Ramsey –en particular, en la ardua cuestión de en qué medida el sistema secundario aporta contenido a la teoría que no hubiera quedado ya recogido en el sistema primario. El celoso escrutinio de estas disputas –veremos– hace pensar que obra *velis nolis* en el funcionalismo, y por ende en la psicología cognitiva, un ineluctable compromiso con la idea de que el vocabulario teórico de una psicología científica madura incluirá términos referidos a estados mentales no sólo a modo de *definiendum* sino también de *definiens*.

Entre los réditos que auspiciaba pensar en la mente a la luz de la teoría de autómatas, administrando además escrupulosamente la distinción entre identidad de tipos e identidad de casos, resultaba particularmente estimulante la expectativa de poder dotar a los estados mentales del vigor causal que su análisis disposicional le denegaba. De ese peso que cada estado mental devengaba como causa de la conducta o de otros estados mentales en tanto que era idéntico a un estado neurofisiológico, entretejido con el hecho de que la taxonomía de lo mental a la que habríamos de asirlo no se construiría bajo criterios neurofisiológicos, sino funcionales, destilaba no en vano la perspectiva de que la psicología pudiera contar con un nivel autónomo de explicación en el edificio de la ciencia. **Eficacia causal, relevancia explicativa, autonomía**: tales son, así pues, los polos entre los que clareaba un debate todavía inconcluso.

Aun atendidas las objeciones de Wittgenstein en cuanto a que el comportamiento no puede ser efecto de un proceso mental toda vez que es parte del concepto de dicho proceso, y conjurada la advertencia de que entenderlo así conlleva incurrir en la metáfora paramecánica denunciada por Ryle, la polémica dista, en efecto, de poder darse por cerrada. Porque si la *eficacia* causal que el funcionalismo puede reconocerle a los estados mentales se restringe a la que cada estado mental atesore en virtud de las propiedades nerviosas –bioquímicas… físicas– del estado neurofisiológico en que venga encarnado, si, dicho de otro modo, nunca son esas propiedades psicológicas según las cuales lo clasificamos al lado de otros estados mentales las que lo hacen trabarse en una cadena de causas y efectos, no es descabellado argumentar entonces que el anhelo de *autonomía* epistemológica para la psicología, si por tal cosa se entiende su reconocimiento como explicación soberana, última, de ciertos ámbitos de la realidad, acabará disolviéndose en la mera concesión de alguna forma de *relevancia* de las teorías psicológicas a efectos pragmáticos, ya sea como un lenguaje burdo pero en ocasiones provechoso en el que condensar regularidades cuya explicación bien podríamos detallar, más pausadamente, en términos fisiológicos, ya como un mero bálsamo ante nuestra transitoria ignorancia de dichos detalles. La crítica de los lúcidos argumentos de Frank Jackson y David Braddon-Mitchell, quienes ven una quimera en toda idea de autonomía que desborde los márgenes de esa relevancia pragmática, nos permitirá cartografiar una travesía –la que suelta amarras en la distinción entre identidad de casos e identidad de tipos y trata de enrumbarse a la irreductibilidad de las teorías psicológicas– que muchos han tachado de impracticable, y en la que se ha avistado a veces, como en el

problema del retorno de lo mental, los escollos entre los que la herencia del conductismo habría condenado a la psicología cognitiva a gobernarse.

La construcción de autómatas que remedasen el comportamiento de los seres vivos es, desde luego, un empeño que antecede con mucho al desarrollo de la noción de máquina abstracta que llegaría de la mano de Turing, o a la reflexión de Putnam sobre sus implicaciones de cara a nuestra comprensión de la naturaleza de la mente. ***Ab Architae columba lignea*** se remonta la **madrugada del autómata**: desde que una paloma de madera armada por Arquitas de Tarento surcara el cielo de la Magna Grecia hasta que el célebre *Canard Digérateur* de Jacques de Vaucanson asombrase a las cortes europeas, la exactitud de la imitación venía siendo el mayor orgullo del artífice. Pero una vez que esa pequeña vanidad cediera a la industrialización, los afanes del constructor de autómatas pronto abandonarían el ámbito de la mímesis para ceñirse a un designio aún vagamente aprehendido de reproducir los *principios subyacentes* o los *rasgos esenciales* del fenómeno biológico o psicológico en cuestión. En el esfuerzo por entender qué era exactamente lo que debían compartir el autómata y el organismo cuyo comportamiento se trataba de reproducir comenzó a vislumbrarse la idea de lo mental que cobraría carta de naturaleza en Putnam. Los primeros autómatas fototrópicos, de hecho, alumbrarían –nunca mejor dicho– la temprana conclusión de Herbert S. Jennings, en el seno de la polémica sobre la naturaleza de los tropismos que mantenía con Jacques Loeb, de que existen principios generales relativos a la conducta de los distintos organismos, así como de ciertas máquinas, que desbordan aquello de lo que pueden dar cuenta los recursos de la física y la química. A los ingenieros dedicados a la construcción de autómatas, además, no podía escapárseles el hecho de que lo que quiera que hubiesen logrado merced a un dispositivo, pongamos por caso, eléctrico podía sin duda replicarse con uno mecánico, o magnético, o hidráulico, o químico… para Silas Bent Russell, Thomas Ross, William Grey Walter, Herbert Edgard Coburn, Anthony G. Oettinger o J. Anthony Deutsch era enteramente transparente que la particular estructura física, ya fuera orgánica o inorgánica, en la que se implementara cierta capacidad no podía ser la viga maestra de la comprensión de dicha capacidad. Más allá de la equivalencia conductual por la que solían desvelarse los ingenieros –y que tanto Clark L. Hull como Alan Turing, al proponer su célebre juego de imitación, habían dado como criterio válido para concluir que un autómata muestra idénticas capacidades que el organismo que trata de replicar–, sería Kenneth J.W. Craik, un filósofo y psicólogo formado en Edimburgo y Cambridge, quien sabría entrever que se tornaba imprescindible un análisis abstracto de la tarea que permitiera articular criterios de equivalencia más exigentes. No pocos conductistas, capitaneados por Clark L. Hull, se interesarían también por lo que Hull bautizó como el *enfoque del robot*: John M. Stephens había logrado una simulación electromecánica de la ley del efecto, Harry D. Baerstein, Robert G. Krueger y el propio Hull tratarían de construir un artefacto sensible al condicionamiento pavloviano que mostrara también sus diversos fenómenos asociados, como la extinción –lo mismo que intentaría hacer Lewis B. Wyckoff para el condicionamiento operante… El paulatino desencanto de Hull con

aquellos proyectos, quizá acrecentado por el desinterés o el rechazo de otros conductistas, acabaría relegándolos, cuando no al olvido, a un rincón apartado de las preocupaciones de Hull. Que Hull, como Tolman, se había anticipado al incipiente cognitivismo lo entendería sin titubeos Skinner al repudiar taxativamente su enfoque: cuando Miller, Galanter y Pribram se aprestasen a defender la simulación como método de investigación psicológica, no les sería difícil invocar el recuerdo de Clark L. Hull y su enfoque del robot.

Hay, quizá, un núcleo de motivos que alentó de manera espuria el auge del cognitivismo, y que tienen que ver con cierta malinterpretación de los primeros resultados de ese estudio abstracto de la naturaleza de los autómatas que propiciara el declive de los afanes groseramente miméticos con que nacieron –en particular, con una lectura mistificada de la tesis de Church-Turing. Cuando Turing, en las seminales investigaciones que verían la luz en 1936 y 1937, se propone abordar una vertiente de ese análisis abstracto de las tareas en el que Craik reclamaba ahondar unos años después, su decidido propósito es delimitar la clase de tareas que pueden ser llevadas a cabo de forma mecánica –siguiendo un *procedimiento efectivo*, un algoritmo. Es decir: la clase de tareas que un operador humano provisto sólo de lápiz y papel puede consumar ayudado de su perseverancia pero no de su ingenio. El *computador* es para Turing, una y otra vez, ese minucioso empleado, a menudo exhausto por el tedio, y su radical propuesta es que las máquinas lógicas por él ideadas son capaces de redimir al computador de su onerosa servidumbre, no muy distinta de la que Marx y Engels habían denunciado que se imponía al obrero. De hecho, *computadores* venían siendo desde antiguo los religiosos dedicados al cálculo de la fecha del Domingo de Gloria, como lo serían después cartógrafos, astrónomos y contables, y eran sus duras condiciones de trabajo, así como los numerosísimos errores que acababan cometiendo, lo que ya había incitado a Charles Babbage a tratar de construir máquinas que pudieran reemplazarlos. Más audaz si cabe, Turing vislumbró la idea de que en realidad su máquina lógica, adecuadamente programada, podía acometer cualquier tarea de esa naturaleza, con lo que ni siquiera sería necesario construir una máquina diferente para cada uno de los quehaceres de los computadores: se trataba, así pues, de una máquina universal. Si la semilla de las preocupaciones de Turing se encontraba en **La extenuación del computador**, el remedio habría de hallarse en el alumbramiento un procedimiento universal de cálculo**: contra *capitis defatigatione, mathesis universalis***.

Ésa era, pues, la *tesis de Turing*: que la noción de tarea mecánica –de procedimiento efectivo, de algoritmo– y la noción de tarea computable mediante las máquinas lógicas por él definidas son equivalentes. El propio Turing demostraría que la reconstrucción del concepto de algoritmo lograda poco antes por Alonzo Church en términos de definibilidad mediante el aparato del cálculo lambda quedaba subsumida en su tesis, que desde entonces conocemos como *tesis de Church-Turing*. Pues bien: es sobre la idea de que la tesis de Church-Turing implica que una máquina de Turing es capaz de resolver *cualquier tarea* –así, sin restricciones, o bien con restricciones sensiblemente más laxas que las fijadas por el propio Turing– sobre

la que pesa la acusación de venir impulsando ilegítimamente la concepción cognitivista de la mente. Escarbar en las raíces de esa interpretación viciada, cuya huella encontraremos en pensadores cuya afinidad con el cognitivismo es bien diversa –Dennett, Churchland y Churchland, Johnson-Laird, Guttenplan, Searle– nos conduce hasta la provocativa lectura del trabajo de Warren S. McCulloch y Walter Pitts sobre la representación matemática de las funciones nerviosas que propusiera John von Neumann pero también, irónicamente, hasta el enardecimiento con que Turing, ya en 1950, daría en defender la *hipótesis* de que una máquina llegaría algún día a superar el *juego de imitación* que él proponía como criterio para reconocerle inteligencia –o mejor, como sustitución de la pregunta por su inteligencia–, es decir, el célebre *test de Turing*. También John McCarthy, Marvin Minsky, Nathan Rochester y Claude Shannon, cuando enunciaron en 1955 la promesa fundacional de la investigación en inteligencia artificial, la perfilaron como la *conjetura* de que existirían máquinas capaces de reproducir cualquier aspecto del aprendizaje o la conducta inteligente humana: de lo que se trata, así pues, es de intentar entender, siquiera por encima, como una conjetura apasionante iría tornándose en poco menos que un presunto teorema. Quizá –veremos– esa mutación viniera en parte avivada, soterradamente, por el antiquísimo ensueño de hallar la *characteristica universalis*, el prodigioso formalismo de cálculo que nos dotase del poder de resolver de forma mecánica cualquier asunto sobre el que la razón pudiera pronunciarse, un anhelo que nos haría remontarnos desde las máquinas de Babbage a las de Leibniz, y de allí a Descartes, a Raimundo Lulio, y a la *anagoge* del prisionero que escapa de la caverna platónica o el alma que sobrevuela la espalda del cielo en *Fedro*.

Las máquinas pensantes y la crisis de fundamentos de las matemáticas que venía sacudiendo el pensamiento europeo desde mediados del s.XIX guardan entre sí lazos tan estrechos como los que anudan ambas historias con el perenne espejismo de la lengua perfecta. De hecho, la "aplicación al *Entscheidungsproblem*" a la que hace referencia el título de los trabajos de Turing de 1936 y 1937 remite directamente a la tercera de las cuestiones con que David Hilbert acuciara a la comunidad matemática en 1928: si las matemáticas son completas, si son consistentes, si son decidibles. El anhelo de Hilbert –un límpido trasunto matemático de la lengua perfecta– era fundar todo el conocimiento matemático en un conjunto finito, completo y consistente de axiomas expresado en un lenguaje formal, de suerte que la verdad o falsedad de cualquier enunciado matemático, adecuadamente formulado en ese formalismo, pudiera dirimirse mediante un procedimiento algorítmico. Ya en 1931, Kurt Gödel había demostrado que consistencia y completud son exigencias incompatibles; las máquinas lógicas servirían a Turing para demostrar, por *reductio*, que la decidibilidad es también una quimera –lo cual, dicho sea de paso, da un aire irónico a la tosca mistificación que, como acabamos de ver, aguardaba a su trabajo. Además, la naturaleza autorreferencial de la prueba armada por Turing evocaba poderosamente tanto el argumento de Gödel como la paradoja de Russell, que Hilbert, de manos de Ernst Zermelo, conocía ya desde los primeros años del siglo.

Nada, sin embargo, haría desistir a Hilbert de su abominación del *Ignorabimus!*, y en su empeño infatigable por reconstruir el método axiomático frente a tanta adversidad veremos florecer una concepción de las teorías como esquemas conceptuales aplicables a infinidad de realidades cuyas huellas –quizá tanto, si Neil Tennant está en lo cierto, como las de la noción abstracta de función desplegada por Lejeune Dirichlet– son nítidas entre quienes, como Putnam, Pylyshyn, Fodor o Johnson-Laird, han tratado de poner en limpio los planos de los cimientos funcionalistas del cognitivismo –incluso en los temores de quienes, como Block, se han preocupado de acotar la mansedumbre de esos esquemas conceptuales para evitar que puedan acabar, como ya Frege denunciara, refiriéndose a casi cualquier cosa, o entre quienes, como Searle, han querido ver en esa mansedumbre una refutación terminante del propio cognitivismo.

La cuestión candente, así pues, es si cualquier sistema que viniera a satisfacer las relaciones entre sus elementos postuladas por cierta teoría habría de ser considerado como un modelo de la teoría: en el caso de la teorización psicológica, si cualquier sistema del que pudiera con verdad predicarse la teoría sería, en consecuencia, una mente. O, dicho de otro modo, si la amable metáfora al amparo de la que tantos psicólogos venían desarrollando su trabajo –la computadora de la mente–, cuyas aristas podían limar sin esfuerzo, no habría de ser tomada, como incansablemente reiteraría Pylyshyn, al pie de la letra, si la asunción del cognitivismo no nos comprometería a admitir también que una computadora debidamente programada *es* una mente, y que la mente *es* una computadora. La constatación de que eludir ese compromiso sea tan común como afrontarlo, desde luego, es menos de lo que necesitaríamos para justificar que lo eludamos –como es menos, también, el hecho de que eludirlo halague alguna vanidosa visión de nosotros mismos que podamos esgrimir mediante la retórica de la usurpación o la deshumanización. No es fácil, en particular, dejar atrás la interpretación literal de la analogía entre mentes y computadoras sin ser arrastrado de vuelta a las escolleras del conductismo, del fisicalismo, o, acaso con más facilidad, de una concepción abierta o veladamente dualista de las relaciones entre mente y cuerpo –así, por ejemplo, veremos como Ulric Neisser, que ya en 1967 se mostraba sumamente crítico con los modelos computacionales de procesos psicológicos al uso, no puede sino refugiarse en la analogía entre lo mental y lo computacional cuando llega la hora de conjurar la perspectiva de una psicología enteramente formulada en el lenguaje de la bioquímica. La razón –piensa Pylyshyn– es que el rechazo de esa interpretación literal proviene de una deficiente comprensión del concepto de computación, que lo liga a los procesos internos que puedan darse en las computadoras que hoy conocemos sin alcanzar a ver la clase general a la que pertenecerían tanto dichos procesos como los procesos mentales.

La idea de que un artefacto mecánico pudiera no sólo imitar la mente de un ser vivo sino llegar a *ser* una mente, como es natural, arraiga en las reflexiones de quienes trabajan en la construcción de tales artefactos mucho antes del advenimiento de la ortodoxia cognitivista en psicología; entrelazada con ella va madurando

también poco a poco la noción de que el programa -el diseño, el croquis– de la máquina en cuestión pueda constituir ni más ni menos que una teoría de los procesos mentales que se haya logrado remedar. **Mentes y máquinas**, entonces, ligadas entre sí por un vínculo cuya naturaleza iríamos tratando de acotar por medio de distintas metáforas –**metáforas de una metáfora**–, para recabar al fin en la controvertida constatación de su literalidad. Así, leíamos ya en Bent Rusell que la máquina encarna una hipótesis teórica, en Loeb que cierta teoría es la base de la construcción de tal o cual máquina, en Stephens que la máquina es una explicación mecánica de la teoría; Hull verá esos extraños artefactos que iban colonizando los laboratorios de psicología como hipótesis mecánicas y Tolman como ejemplificaciones de teorías; Wallace describirá cierta máquina como una contribución a una teoría, Deutsch como un modelo que la encarna, Wyckoff como un instrumento que permite su contrastación empírica, Grey Walter como una hipótesis cristalizada, MacKay como una plantilla construida según cierta hipótesis de modo que podamos confrontarla con la realidad… pronto, Coburn se atrevería ya a plantear que el croquis del aparato podría constituir en sí mismo tanto una teoría como una herramienta de investigación, mientras Broadbent advertía que una teoría puede expresarse en palabras o ecuaciones tanto como en los componentes materiales de un ingenio mecánico. Cuando en 1960 Miller, Galanter y Pribram apuntan que el programa de ordenador que imita un proceso psicológico constituye una teoría de ese proceso tan aceptable como la ecuación que lo describe –qué duda cabe– las raíces de su dictamen son ya profundas.

Al mismo tiempo, como sabemos, el repudio de la mera *imitatio* de la naturaleza de la que hacían gala los viejos autómatas iría paulatinamente dando fuerza a una cuestión crucial: la de delimitar cuáles son las similitudes relevantes y cuáles las despreciables a la hora de decidir si la teoría de determinada facultad psicológica encarnada en la máquina es correcta, que sería tanto como decidir si la máquina posee de hecho dicha facultad. Que el comportamiento de la máquina resultara, para un juez experto, indiscernible del humano en el transcurso de una conversación –lo esencial del juego de imitación imaginado por Turing– serviría de punto de partida, pero las exigencias no tardarían en endurecerse cuando Newell, Shaw y Simon se aprestaran a la comparación de trazos del registro de sus computadoras con protocolos introspectivos humanos, poniendo así en práctica una exigencia de identidad de procesos internos a la que ya habían apuntado, de una manera o de otra, Needham, Ellson, Bradner, Ross, Pask o Grey Walter, y que Fodor consagraría como criterio de equivalencia funcional fuerte junto con la equivalencia no sólo de las conductas exhibidas sino también de las posibles –es decir, equivalencia no sólo en el ámbito de la actuación sino también en el de la competencia. En el propio trabajo de Turing encontraría luego Putnam una forma elegante de expresar esa idea: la identidad entre tablas de máquina. Lo importante, en todo caso, era otra cosa: se había abierto la pregunta por el nivel de abstracción adecuado para la comprensión de lo mental.

Antes de ahondar en el estudio de cómo todo esto habría de alterar nuestra comprensión de las relaciones entre los procesos mentales y los cerebrales –y, en particular, de si deja o no espacio para una noción de eficacia causal de lo mental que pudiera otorgar a la explicación psicológica cierto grado de soberanía respecto de la explicación neurofisiológica que no manara de la indigencia de nuestro conocimiento del sistema nervioso–, dedicaremos un breve **Interludio** a sopesar una de las críticas más feroces de la que ha sido objeto la psicología cognitiva: la que la perfila como una manifestación de la misma ideología deshumanizadora que subyace a la casi inconcebible brutalidad desencadenada durante la Segunda Guerra Mundial y a su principal detonante, el auge del nazismo. La metáfora del burócrata –que hemos visto germinar en Turing y de la que no nos será difícil encontrar ramificaciones en las propuestas de explicación homuncular de Dennett, en la noción de jerarquías cognitivas de Minsky, en Newell y Simon o en Pyslyshyn– reflejaría a ojos de John Shotter la misma concepción de lo humano que opera en la ideología tecnocrática que ya en 1922 denunciara Max Weber, y que hizo posible la existencia de los campos de exterminio. La ambición naturalista del cognitivismo no sería sino el núcleo de esa constelación de ideas deshumanizadoras. Espigaremos, sin embargo, algunos ejemplos de cómo la metáfora burocrática ha acompañado a la mecanización de la imagen del mundo desde los mismísimos orígenes de la *nuova scienza* galileana –no en vano puede hallarse ya en la obra de Johannes Kepler–, y desde luego desempeñó un papel relevante en el desarrollo de conductismo; veremos, además, cómo la psicología cognitiva ha querido entenderse a sí misma como un movimiento emancipador, que buscaba salvaguardar de la hoguera del conductismo, o del eliminacionismo, ciertos valores humanistas y que, en lo relativo a la concepción de lo mental bajo la metáfora burocrática, pasaba precisamente por la flexibilización de buena parte de las jerarquías de control que regirían los procesos psicológicos. Si bien ninguna de estas observaciones serviría para dar por refutados los argumentos de Shotter, en ellas podrían atisbarse matices de cierta relevancia de cara a un análisis más detenido de lo que su énfasis en **Autómatas y oficinistas** entraña para **el cognitivismo como ideología.**

Ha quedado ya bosquejada la conclusión de que, de la mano de las reflexiones sobre aquellas asombrosas máquinas capaces de remedar comportamientos humanos que creíamos mediados por procesos psicológicos, o, al amparo de la investigación sobre los fundamentos de la matemática, sobre la naturaleza abstracta de algunos de esos procesos, fue sin duda la crítica de la concepción fisicalista de lo mental que comenzaba a hacerse fuerte sobre las ruinas del conductismo lo que alentó el florecimiento del funcionalismo como sustrato de la nueva psicología cognitiva. El corazón de esa crítica es la idea de **La mudable encarnación de lo mental** que quedara desplegada en la investigación sobre "La naturaleza de los estados mentales" que Putnam dio a la imprenta en 1967, cuatro años después de haber publicado sus argumentos modales contra el conductismo. Pero advertiremos, al mirar en detalle, cierta asimetría en aquellos razonamientos de Putnam: mientras las objeciones contra el fisicalismo, como sabemos, descansan sobre su caracterización

como una tesis descabelladamente ambiciosa y, por tanto, de una vulnerabilidad empírica desmedida, los argumentos que habrían de convencernos, una vez abandonado el fisicalismo, de abrazar el funcionalismo aluden más bien a cuestiones de índole conceptual, que tienen que ver con nuestra manera habitual de identificar estados mentales en nosotros mismos o en los demás. Ahora bien: comoquiera que es la presencia de ciertos comportamientos ante ciertas circunstancias lo que parece operar al menos en ese trámite ordinario de atribución a otros de estados mentales análogos a los nuestros, el argumento de Putnam corre el riesgo de deslizarse de nuevo hacia un análisis conductual del vocabulario mentalista, riesgo que Putnam trata de conjurar –acaso sin confiar del todo en que la credibilidad del conductismo haya quedado irremediablemente minada por sus ejercicios de razonamiento modal sobre el dolor– por la vía de interponer una distinción entre nuestros conceptos psicológicos y las propiedades a las que aluden, distinción cuyos trazos el fisicalista avisado podría emplear para debilitar la posición del propio Putnam. Un cierto equilibrio precario, en suma, parece obligado reconocer en el armazón argumentativo sobre el que se erigirá la ortodoxia funcionalista. La insistencia de Block en aclarar que lo que deja desabrigado al fisicalismo no es el mero hecho de que estados mentales del mismo tipo puedan cristalizar como estados cerebrales heterogéneos, sino la inabarcable dificultad de concebir siquiera propiedades de primer orden –neurológicas, si se quiere, o físicas– que diluyan esa heterogeneidad, no termina de deshacer aquella inestabilidad originaria. No es raro, así pues, que las conclusiones antirreduccionistas que parten del trabajo de Putnam se hayan visto sometidas a un cuestionamiento tan enérgico como minucioso entre quienes –no siempre, tal vez, plenamente conscientes de la envergadura de la tarea– aspiran pese a todo a una implacable reducción del conocimiento psicológico al ámbito teórico de las ciencias del sistema nervioso.

La controversia en torno a la interpretación de las tesis funcionalistas, tanto en términos epistemológicos como ontológicos, parece brotar de los mismos manantiales que la propia concepción funcionalista de lo mental, y es tan temprana como ésta. De hecho, es atendiendo a esa controversia como acaso con mayor fidelidad a su despliegue histórico cabe trazar –no ya del funcionalismo, sino, por hacer honor a la pluralidad de matices que la vertebra, de los **Funcionalismos**– una **cartografía teórica**. Así, nos hallaríamos ante un funcionalismo de corte analítico cuando las generalizaciones acerca de los lazos que guarde un determinado tipo de estado mental con estímulos, respuestas y otros estados mentales provengan del estudio conceptual de la psicología ordinaria, o de sentido común, y con un funcionalismo de corte empírico cuando la fuente de tales generalizaciones se haga residir en teorías desarrolladas en el seno de la psicología científica. El funcionalismo analítico se perfila, entonces, como una decidida apuesta por la confluencia entre los roles funcionales aislados como resultado de ese examen de nuestros conceptos psicológicos coloquiales y los roles funcionales desempeñados por los tipos de estados descubiertos en la investigación en fisiología nerviosa –*i.e.*: los ocupantes de esos roles–: es, así visto, una variedad de la tesis de identidad psicofísica, y hereda de

ella su compromiso con el reduccionismo. Algunas de las dificultades que afligen al funcionalismo analítico, como veremos, quedaron desglosadas en un trabajo de Block que era apenas un poco más clemente con el propio funcionalismo empírico. En pocas palabras, es dañina para el funcionalismo analítico su profunda dependencia del concepto de lugar común de la psicología cotidiana, que le impide responder a casos en que ésta no tiene recursos para diferenciar entre estados psicológicos en cuya diversidad nos hace reparar la investigación empírica tanto como a casos en que ésta nos ofrece arrojadas afirmaciones que la investigación revela falsas, o a aquellos otros en los que, por demasiado inusitados, apenas tiene algo que decir.

El funcionalismo empírico, en cambio, germinaría en terrenos netamente antifisicalistas. Si es acertado el esquema de las relaciones entre los conceptos de rol funcional, ocupante de dicho rol, propiedad funcional, instancias y tipos de estados trabado por Stephen R. Schiffer, el sustrato de la interpretación del funcionalismo como una tesis contraria al reduccionismo radicaría en el hecho de que se aísla cada tipo de estado psicológico –digamos, la creencia de que *p*– ligándolo a un rol funcional, delimitado en el trabajo empírico de los psicólogos, tal que la propiedad de ser una instancia del tipo de estados caracterizados por poseer dicho rol funcional es idéntica a la propiedad de ser una instancia de dicho tipo de estado psicológico: pertenecer a un tipo de estado psicológico es, así pues, una propiedad funcional –es decir, la propiedad de pertenecer a un tipo de estado con determinado rol funcional–, y, nada impide que las propiedades físicas de las diversas instancias que forman un tipo de estados, agrupadas en tanto que comparten un determinado rol funcional, puedan ser heterogéneas. Pero en el funcionalismo, si todo esto es así, habría –*pace* Block– algo más que una explicación funcional del comportamiento de un organismo o sistema construida en términos de estados internos que se identifican por sus relaciones con estímulos, respuestas y otros estados internos: formaría parte inextricable del funcionalismo una teoría acerca de la naturaleza de las actitudes proposicionales en la que la proposición actúa como un índice para el rol funcional de un tipo de estados internos cuyas instanciaciones poseen la propiedad funcional que identifica a una determinada actitud proposicional. Ésa es, en el fondo, la misma conclusión realista acerca de las representaciones mentales a la que, por distintos derroteros, había arribado ya Fodor en su influyente estudio sobre *La explicación psicológica*.

Es habitual diferenciar en el seno del funcionalismo empírico, de la mano de William Bechtel, entre un primitivo funcionalismo de tabla de máquina, un más refinado funcionalismo computacional, y un tardío y controvertido funcionalismo homuncular, ligado a veces a una concepción instrumentalista de los estados mentales. Tras discutir, de la mano de Block, si sobre lo que todas las variantes del funcionalismo empírico comparten –a saber: la idea de que nuestra noción de lo mental ha de venir dictada por el desarrollo científico de la psicología– es factible construir un argumento *a priori* en su favor, habrá que adentrarse, primero, en los motivos que llevaron a Block y Fodor, ya en 1972, a distanciarse de las primeras propuestas de Putnam, y, luego, a otros pensadores de la órbita funcionalista a

buscar inspiración en la concepción instrumentalista de lo mental desarrollada por Daniel C. Dennett para tratar de esquivar algunas de las dificultades que se adivinaban en el funcionalismo computacional de Block y Fodor, dando alas así al funcionalismo homuncular.

La clave del primer tránsito, como veremos, reside en el convencimiento de que una concepción netamente funcionalista de la mente puede tenerse en pie aun renunciando a la ambición de dar a cada sistema capaz de albergar estados mentales una única descripción formal como autómata probabilista de suerte que cada tipo de estado mental del sistema en cuestión quede identificado con un estado de tabla de máquina del autómata descrito. La cuestión, entonces, era cribar mena y ganga en el yacimiento del audaz proyecto de Putnam, cuyas numerosas flaquezas pronto quedaron a la vista: Block y Fodor advertían ya de las serias dificultades que el funcionalismo de tabla de máquina había de arrostrar a la hora de dar cuenta de la distinción entre estados psicológicos disposicionales y actuales, de la intervención simultánea de varios estados psicológicos en la determinación de la conducta, del carácter cualitativo de ciertos estados mentales, del modo en que la clasificación cotidiana de estados mentales en tipos ignora diferencias irrelevantes, y de la productividad y sistematicidad inherente a muchos estados psicológicos en virtud de su carácter proposicional. Algunas –ciertamente no todas– de esas dificultades quedarían mitigadas, de acuerdo con Block y Fodor, si pensábamos en los estados mentales no como estados de tabla de máquina de un autómata, sino como sus estados computacionales, incorporando así al aparato explicativo de las teorías psicológicas, y a la concepción funcionalista de lo mental que se ofrecía como sustrato suyo, el concepto nuclear de computación sobre representaciones: construir una teoría psicológica equivaldría, en suma, a detallar las reglas de manipulación y las representaciones que intervienen en el control de la conducta de determinados organismos o sistemas. Con ello, por otra parte, quedaban derogados los criterios de equivalencia psicológica inspirados en el juego de imitación de Turing y comenzaba a asentarse una noción de equivalencia más rigurosa, que exigiría no sólo la indistinguibilidad de las respuestas sino también la de los procesos internos –es decir, la de las representaciones y las computaciones involucradas en la producción de la respuesta–; en el seno de esa noción estricta de equivalencia veremos fructificar las nociones de algoritmo y de arquitectura funcional.

En cuanto al segundo trayecto –el que conduce del funcionalismo computacional al funcionalismo homuncular–, el giro crucial habría radicado en advertir el papel decisivo que desempeña en el desarrollo de modelos de simulación cognitiva la descomposición de cada tarea, o proceso, en otras más y más sencillas, hasta que el carácter propiamente cognitivo de la tarea haya quedado diluido –es decir, hasta que se trate ya de una tarea mecánica en el sentido de Turing: que no requiera el concurso de la inteligencia, redimiendo el préstamo explicativo que habíamos tomado al postular sucesivos homúnculos a cargo de cada una de las tareas más complejas. Prestar atención, entonces, a las estructuras jerárquicas habituales en el ámbito de la inteligencia artificial, a los diagramas de flujo y los

árboles de subrutinas, se perfilaba como la forma de armar de una vez por todas una concepción de lo mental que pudiera vencer las trabas que surgían a cada paso. Lo que aprenderíamos al hacerlo, en pocas palabras, es que adjudicar estados o procesos psicológicos al sistema en su conjunto –el organismo, el autómata…– no tiene ni más ni menos justificación que atribuírselos a cada uno de los homúnculos que intervienen en el modelo o la teoría –a cada uno, claro, a tenor de sus capacidades. El funcionalismo homuncular y un instrumentalismo acerca de lo mental a menudo rayano en el eliminacionismo quedaban así estrechamente coaligados; con ello, se quebraba el compromiso con una concepción realista de la mente que bajo la batuta de Fodor el funcionalismo computacional venía por lo general considerando irrenunciable, y se abocaba al funcionalismo a enarbolar una visión de lo mental sensiblemente más apartada de las intuiciones del sentido común. A cambio, si los paladines de la idea del homúnculo están en lo cierto, la suavización de los límites entre las nociones de estructura y función, así como la posibilidad de incorporar conceptos teleológicos al análisis funcionalista de la intencionalidad, ensancharían el horizonte de nuestra concepción de la mente y sus relaciones con el cuerpo y el mundo que habita mucho más allá de lo que permitía el aparato explicativo del funcionalismo computacional –apenas podremos dedicar unas apresuradas líneas a sopesar con el mayor cuidado esas reivindicaciones.

La ruptura del funcionalismo homuncular con el realismo acerca de lo mental que había imperado hasta entonces en los predios funcionalistas nos ha emplazado ya a mirar cara a cara al hecho de que las venerables nociones de **Espíritu, materia, función**, etc. vienen siendo sometidas a un intenso trabajo de reelaboración: sería hora, pues, de abordar, las diversas **Lecturas ontológicas del funcionalismo** que han tratado de administrar dicho trabajo. Quizá hallemos un buen punto de partida en la constatación de que las tesis funcionalistas guardan una escrupulosa neutralidad ontológica: que la taxonomía natural de los estados mentales deba atender a los patrones de relaciones funcionales que traben tan bien puede conciliarse con la convicción de que cada uno de ellos sea en última instancia un estado físico como con la de que sea en realidad un estado de índole espiritual. Por el contrario, una doctrina materialista que ataña a las propiedades según las cuales los estados mentales se agrupan en clases, restringiendo dichas propiedades al ámbito de las propiedades físicas, colisionaría tan violentamente con el funcionalismo como una doctrina dualista análoga, de acuerdo con la cual hubieran de ser propiedades espirituales las que dirimieran la clasificación de los estados mentales en tipos. En los términos en los que Block ha preferido plantear la cuestión, podríamos convenir, pues, en que el funcionalismo implica rotundos compromisos metafísicos, pero ningún compromiso ontológico. Sin embargo, conviven en el seno del funcionalismo –como ya se ha ido entreviendo– el convencimiento de que éste no supone una quiebra con los principios del fisicalismo –sino que, antes bien, constituye la interpretación más coherente de una tesis de identidad psicofísica con alcance de tipos– con el convencimiento contrario de que funcionalismo y identidad psicofísica de tipos son abierta e irremediablemente discordantes. Si esta última es la posición

que, en la estela de Putnam, habría defendido Fodor –tesis de identidad de estados funcionales, según la denominación acuñada por Block–, la primera –es decir, la tesis de identidad de especificaciones funcionales–, habría encontrado en Armstrong, bajo el auspicio de Lewis, a su paladín; si una entronca, pues, con la interpretación empírica del funcionalismo, la otra lo hace con su interpretación analítica.

Una vez más, la controversia puede cifrarse en un litigio sobre el concepto de dolor, que Lewis, en su reseña del trabajo sobre "Psychological Predicates" que Putnam había publicado en 1967 –luego rebautizado como "The Nature of Mental States"–, asemejaba al de número premiado en un sorteo, con el propósito de hacer ver que el hecho de que un mismo estado mental, como el dolor, pudiera identificarse unas veces con un cierto estado cerebral y otras veces con otro no tenía nada de particular –no más, al menos, que el hecho de que el número premiado en un sorteo pudiera ser unas veces uno y otras veces otro. Conviene apuntar, no obstante, al margen de las reflexiones de Lewis, que de esto se deriva que la tesis de que el dolor es un estado cerebral ha de entenderse en el mismo sentido en el que entenderíamos la de que la propiedad de ser el número premiado en un sorteo es una propiedad matemática, entonces el fisicalismo queda, como poco, severamente vaciado de contenido. Además, a la contundente censura del prejuicio según el cual los nombres de los estados mentales deben siempre referirse a dichos estados mentales de forma necesaria e independiente del contexto, tan acertadamente articulada por Lewis, cabe enfrentar una reprobación igualmente firme del prejuicio, que cabe imputarle, según el cual los nombres de los estados mentales no pueden nunca referirse a dichos estados mentales de forma necesaria e independiente del contexto por la sencilla razón de que existan nombres de estados cerebrales que no lo hacen y con los que, contingentemente y en ciertos contextos, comparten referencia. Por último, acaso no esté de más recordar las objeciones que el propio Lewis presentaría años después, en el marco de su defensa de la posible existencia de estados de dolor funcionalmente atípicos –el "dolor insensato"–, al intento de construir un concepto funcional de dolor en términos disyuntivos, que tildaría de "estrategia desesperada": es difícil ver por qué idéntica renuencia no haya de oponerse a la construcción de un concepto neurofisiológico de dolor en términos disyuntivos. Con todo, no deja de ser verdad que la disputa acaba estancándose en la brecha que el propio Putnam dejara abierta al afilar su crítica del conductismo lógico: bien puede Lewis argumentar que todo eso concierne a fin de cuentas a nuestro concepto de dolor, pero resulta inocua respecto a la naturaleza del dolor considerada al margen de nuestros conceptos. Así vista, la controversia entre Lewis y Putnam queda destilada en si hemos de concluir, con Lewis, que cada tipo de estado mental se identifica con un tipo de estado físico que ocupa un determinado rol funcional, o más bien, con Putnam, que se identifica con un conjunto posiblemente dispar de estados físicos que comparten la propiedad funcional de ocupar el rol característico de ese tipo de estado mental –es decir, con un tipo de estado funcional.

En cualquier caso, la idea de neutralidad ontológica que aletea en la lectura del funcionalismo desplegada por Block hunde sus raíces, además de en el trabajo de

Putnam o Fodor, en la propuesta de análisis temáticamente neutral de los enunciados que usamos como informes de nuestras sensaciones que ya en 1959 había perfilado Smart. Poco antes, la insistencia de Place en que la tesis de identidad psicofísica atañía a las relaciones entre estados mentales y estados del sistema nervioso central, no entre descripciones psicológicas y descripciones neurofisiológicas de dichos estados, había dado paso al planteamiento de dicha tesis apelando a la distinción fregeana entre *Sinn* y *Bedeutung*, que germinó en el pensamiento de Herbert Feigl. Pero la viabilidad de la táctica de Feigl era muy discutida: no en vano, Arthur Prior había usado la distinción entre connotación y denotación forjada por Mill, nítidamente cercana a la de Frege, precisamente en el marco de un intento de afinar las críticas de George E. Moore a la falacia naturalista en el ámbito de la ética, del que concluía que sólo desproveyéndolas de toda pretensión analítica –entendiéndolas, esto es, como una tesis netamente empírica– podía el naturalista hacer sostenibles sus tesis. Sería Max Black quien advirtiera la fricción entre el recurso a Frege de Feigl y el recurso a Mill de Prior e hiciera ver que el hecho de que el concepto de dolor y el del estado cerebral con el que el dolor quedara identificado no compartieran el mismo sentido fregeano entrañaba que el dolor y tal estado cerebral difiriesen en al menos una propiedad, su modo de presentación, por lo que aun concediendo que la universal coincidencia de uno y otro nos abocara a la conclusión de que el dolor es en efecto aquel estado cerebral, tendríamos que seguir atribuyendo al dolor una naturaleza inexpugnablemente mental, ligada a su peculiar modo de presentación –el dolor sentido. La respuesta de Smart al vigoroso desafío de Black sería, como veremos, el análisis temáticamente neutral: la reformulación de nuestros informes experienciales en un alambicado lenguaje no mentalista como paso previo a la identificación del referente de dichos informes, ya reformulados, con determinados estados cerebrales. Las objeciones no se harían esperar: entre las más gravosas para Smart se cuentan las blandidas por James W. Cornman, quien apuntaba que los informes temáticamente neutrales quebraban las relaciones de transitividad que cabría esperar si fueran sinónimos de los informes experienciales, y que recuperarlas entrañaba por fuerza renunciar a la pretendida neutralidad. Los argumentos con los que Block refutara la idea de que el funcionalismo sea una mera variante del conductismo lógico pueden, por otra parte, plegarse también a la tarea de rebatir la misma idea respecto de las relaciones entre funcionalismo y fisicalismo, que es el núcleo de la tesis de Lewis. Con un gesto ya familiar, Fodor apelaría sin dudarlo a la cosecha de generalizaciones que hubiera de ofrecernos una taxonomía funcional de los estados mentales como argumento epistemológico decisivo para entender el funcionalismo bajo un prisma antirreducionista, pues la interpretación fisicalista nos forzaría a tomar la vigencia de esas generalizaciones como hechos brutos.

Al cabo de este nuevo escrutinio de la controversia sobre los lazos entre funcionalismo y fisicalismo, cabe condensar unas conclusiones provisionales en torno el equilibrado balance que esboza Janet Levin: en favor de las tesis de Putnam o Fodor pesa la elegancia con la que pueden manejar la heterogeneidad de estados físicos en que puedan encarnarse ciertos estados mentales; en favor de las de Lewis o

Armstrong ha de contarse, en cambio, la destreza con la que logra vérselas con la cuestión de la eficacia causal de los estados mentales. A las dificultades que afronta un funcionalismo de corte fisicalista para manejar dicha heterogeneidad, ahora bien, han de añadírsele cuantas tienen que ver la radicalización del alcance de la realizabilidad múltiple de lo mental que tiene lugar cuando ésta se aplica no ya a distintas especies o tipos de sistemas físicos, sino a distintos organismos de la misma especie o incluso al mismo organismo en distintos momentos, y que abocaría a ese funcionalista fiel al fisicalismo a una inexorable renuncia al carácter nomotético de sus explicaciones. Ya en 1972 aprestaron Block y Fodor un recuento de las debilidades de la estrategia disyuntiva adoptada por Lewis: la mera suposición de que existan disyunciones físicas distintivas para estados mentales funcionalmente distintos, la presencia de los mismos tipos de estados físicos en las disyunciones con las que se pretendieran identificar distintos estados mentales, contando en cada una de ellas como condición suficiente para alberga el estado mental correspondiente y, por tanto, forzándonos a concluir que el organismo que se encuentra en tal o cual estado físico se encuentra en todos y cada uno de los estados mentales en cuyas disyunciones figure, y, por último, el carácter accidental, no legaliforme, de las disyunciones en cuestión. En ese recuento –veremos– la sombra de esa noción radical de realizabilidad múltiple aparecía velada aquí y allá, pero delinear claramente sus contornos agrava seriamente las dificultades de la interpretación reduccionista del funcionalismo. Ahora bien: no menos severas son las dificultades con que se topa la concepción funcionalista de la mente cuando, habiéndose vinculado a una posición antirreducionista, debe dar cuenta de la eficacia causal propia de los estados mentales –aquella que sustenta las generalizaciones que una taxonomía funcional pueda espigar y que en cambio quedarían fuera de nuestro alcance si forzamos a esa taxonomía funcional a casar con una establecida en el lenguaje de la neurofisiología; tampoco es la primera vez, ni será la última, que el camino nos deja en este atolladero.

El afán por erigir una implacable reducción del conocimiento psicológico al ámbito teórico de las ciencias del sistema nervioso tiene, pues, como labor inacabada e ineludible, la de domeñar esa terca heterogeneidad de lo mental que habría dado en desafiar, insolentemente, la supremacía de la física –o de la fisiología, entendida como su vástago–: acabar, en suma, con **Proteo también encadenado**. Los pujantes **nuevos esfuerzos por la unificación de la ciencia** que han prosperado en los últimos tiempos muestran, junto a la vertiente empírica en la que se encauzan incesantes hallazgos neurofisiológicos, una vertiente conceptual cuyo afluente más caudaloso es acaso el pensamiento de Jaegwon Kim. En su ensayo de refutación de la interpretación antirreduccionista de la tesis de realizabilidad múltiple articulada por Putnam, Kim toma como punto de partida un principio de herencia causal que a su entender es inherente a la relación de instanciación, o realización: las propiedades causales de las clases de las que se predica dicha relación son idénticas. Este principio, como han apuntado Wilson y Craven, se deriva de una intuición que el funcionalismo parece compartir con el fisicalismo –que las propiedades psicológicas

y las propiedades fisiológicas son al fin y al cabo, en cada caso dado, propiedades del mismo fragmento de materia, pero aboca al funcionalista, si Kim está en lo cierto, a un dilema cuyos cuernos son el reduccionismo y el eliminacionismo: no existe, pues, un territorio antirreduccionista en el que buscar abrigo. En la formulación de dicho dilema veremos desplegarse, además de las nociones de clase natural, causalidad e identidad, las de coextensividad nómica y proyectabilidad; en la réplica funcionalista, enarbolada por Block, tomará cuerpo un distingo entre la idea de que las propiedades implicadas en la definición de las clases que damos por coextensivas sean proyectables en un sentido fundado en la noción de justificación, que se rechaza, y la de que lo sean en un sentido fundado en la noción de evidencia objetiva, que se admite, pero que deja expuesto el argumento de Kim a la severa reprobación del fisicalismo que tiempo atrás ensayara Boyd al advertir de cómo el compromiso con una concepción lockeana de las clases naturales como esencias nominales que habían adquirido Place o Smart al defender su tesis de identidad psicofísica de las objeciones basadas en la existencia, contra la ley de Leibniz, de propiedades predicables sólo de uno de los términos de la supuesta relación de identidad resultaba, bajo una mirada más atenta, gravemente lesivo para el fisicalismo que pretendían salvaguardar.

En suspenso, pues, el intento de desautorizar la viabilidad de un funcionalismo ajeno al reduccionismo mediante herramientas conceptuales, queda tantear el expediente de debilitarlo como hipótesis empírica. Si el aluvión de descubrimientos sobre correlaciones entre hechos psicológicos y neurofisiológicos pareciera dar la razón a Putnam al resistirse a traspasar las lindes entre distintas especies, aún podríamos contentarnos, de la mano de Lewis o el propio Kim, con acotar nuestros esquemas de reducción psicofisiológica al ámbito de cada especie estudiada, y argumentar, como ha hecho también Rabossi, que no hay en ello nada que aliente variedad alguna de antirreduccionismo. Las discrepancias en torno a esta estrategia, que ya en 1972 impugnaran Block y Fodor, se han deslizado con los años hacia la cuestión de si existen propiedades físicas –el caballo de batalla en esa lid ha venido siendo la temperatura– cuya instanciación última haya de entenderse como heterogénea en el mismo sentido en el que pueda serlo la de las propiedades psicológicas, sin que ello justifique una lectura antirreduccionista de nuestras teorías acerca de su naturaleza. Pero ese debate –veremos– es estéril, pues aparte de velar a duras penas una falacia *ad auctoritatem*, sólo puede devolvernos a paisajes ya conocidos: en manos del funcionalista queda asumir, *arguendo*, que la temperatura es una propiedad que verifica la tesis de realizabilidad múltiple y que, por tanto, hemos a fin de cuentas de entenderla como un concepto funcional, si es que su uso nos franquea la revelación de regularidades que de otro modo nos eludirían, o, de lo contrario, desecharla como residuo inútil de una teoría obsoleta. Por otra parte, ceder al arrastre de la eliminación sólo en lo que concierne a conceptos psicológicos que pretendan extender su dominio entre especies diversas, escudando la reducción de aquellos que humildemente lo restrinjan a una especie determinada –como, en la estela de Kim, ha planteado Nick Zangwill–, es un pobre cobijo, ya que nos deja

desguarnecidos ante la perspectiva de que encontremos también heterogeneidad en la encarnación fisiológica de estados o procesos mentales del mismo tipo en distintos sujetos de la misma especie, o incluso en el mismo sujeto con el paso del tiempo: nos hallaríamos entonces ante un inventario de conceptos psicológicos llanamente idiográficos, cuya reescritura en el vocabulario de una fisiología de irrenunciable ambición nomotética sería en el mejor de los casos una ociosa acrobacia.

El del posible paralelismo entre **El dolor y la piedra de ijada** –el jade– es un caso práctico en el estudio de las relaciones entre **coextensividad nomológica y herencia causal** en torno al cual ha prendido el debate reciente. Dos rocas metamórficas de estructura química bien diferenciada, la jadeíta y la nefrita, forman la clase natural que llamamos "jade", una clase, por tanto, claramente disyuntiva cuyos miembros muestran sin lugar a dudas propiedades no proyectables. De igual modo que el estudio del jade es infructuoso para la mineralogía, por oposición al de la jadeíta o al de la nefrita, el estudio de estados psicológicos cuya instanciación fisiológica resultara ser disyuntiva –digamos, el dolor– sería terreno yermo para la ciencia de la mente, que sólo podría prosperar –insiste Kim– afanándose en el estudio de los estados fisiológicos en que averiguásemos que el dolor se materializa. Pero la analogía con el jade –contesta Fodor– es espuria: mientras que el concepto de jade consigna una clase formada por la disyunción de las propiedades de la jadeíta y las de la nefrita, provocando así que las propiedades de un miembro dado de dicha clase resulten no proyectables y no nomológicas, los conceptos psicológicos no consignan clases disyuntivas, sino clases en las que se verifica la condición de realizabilidad múltiple: clases cuyos miembros exhiben propiedades proyectables y, en la medida en que son susceptibles de figurar en las leyes propias de las ciencias especiales –de la psicología, en el caso que nos ocupa–, también nomológicas. La razón, según Fodor, es que la identidad entre jade y la disyunción de jadeíta y nefrita es necesaria (es decir: no existen mundos posibles en que otros compuestos distintos de la jadeíta o la nefrita sean jade), mientras que la identidad entre el dolor y el conjunto de condiciones neurofisiológicos que en nuestro mundo lo encarnan es contigente (existen, en otros mundos posibles, variedades de dolor cuya materialización no es ninguna de las que se dan en nuestro mundo). Sin embargo, los argumentos sobre los que Fodor hace reposar esta asimetría modal entre el jade y el dolor son, como veremos, sumamente quebradizos: que nos negásemos a considerar jade una gema sintetizada en un laboratorio que no fuese jadeíta ni nefrita –si es que así fuera–, que admitiéramos, en cambio, considerar dolor el estado interno de un autómata que mostrara el debido patrón de relaciones funcionales –de nuevo, si es que así fuera: darlo por hecho es poco menos que implorar la verdad del funcionalismo, que es lo que se disputa–, o que entre el jade y la disyunción de jadeíta o nefrita se dé la misma relación que entre el agua y el compuesto de dos átomos de hidrógeno por uno de oxígeno son asuntos que atañen dicen más sobre cuestiones pragmáticas, históricas o éticas, o sobre el dispar arraigo de unos y otros conceptos en nuestra cotidianeidad, que sobre una supuesta esencia del jade, o del dolor, que estuviésemos tratando de custodiar. Por otra parte, aunque tanto Kim como Fodor señalan que la idea de que el

dolor carece de propiedades intrínsecas en las que quede sellada su esencia viene a ser un punto de acuerdo entre fisicalismo y funcionalismo, no lo es más que en la medida en que distancia a ambos de una concepción fenomenológica según la cual tal esencia reside en el cariz cualitativo de la vivencia de dolor –en *cómo es* el dolor: no en vano es ese sustrato fenomenológico lo que Lewis pudo esgrimir tanto contra fisicalistas como contra funcionalistas en su conocido *Gedankenexperiment* del dolor marciano y el dolor insensato. Con todo, no será difícil dejar al menos asentada la modesta premisa de que la identidad entre el dolor y su especificación funcional se halla más cerca del núcleo de nuestra concepción del mundo que la identidad entre el dolor y sus encarnaciones fisiológicas –como la identidad entre agua y H_2O que la identidad entre jade y la díada de silicatos. Ni difícil, ni provechoso: esa premisa no puede dirimir la cuestión, pues, como hemos aprendido de Putnam, atañe más a nuestro concepto del dolor que a la naturaleza del dolor.

Un segundo conato de zanjar la controversia ensaya Fodor al ahondar en la distinción entre disyunciones cerradas –jade– y disyunciones abiertas –dolor–, y tratar de elucidar las razones por las que, según su tesis, ni la disyunción ni el concepto disyuntivo consignan propiedades proyectables en las disyunciones cerradas, mientras que, en el caso de una disyunción abierta, el concepto disyuntivo logra consignar propiedades proyectables –múltiplemente realizables–, mientras que la mera disyunción no lo hace. Esas razones –aventura Fodor– tienen que ver con el hecho de que la única aparición conjunta –mejor dicho: disjunta– en leyes científicas de las propiedades que conforman la disyunción abierta ocurre precisamente en las leyes-puente que dan cuenta de su relación con la propiedad de orden superior, o tal vez –dado que eso parece más una reformulación que una explicación– con el hecho de que las leyes que incluyen disyunciones abiertas incitan a pensar en leyes que, apelando a propiedades de orden superior, esquiven la disyunción, y esa incitación brota de la propia naturaleza del razonamiento inductivo.

Tanto Ned Block como Pierre Jacob han bosquejado objeciones a la estrategia de Fodor, que se examinarán detenidamente, y han fraguado sus propias réplicas al desafío de Kim. De los argumentos de Block aprenderemos que la tesis antirreduccionista se sostiene con más firmeza si incorporamos a los motivos conjurados por Fodor para rechazar las disyunciones abiertas el de que cabe exigir que las leyes de la teoría reductora, con el concurso de las leyes-puente, expliquen las leyes de la teoría reducida –además, se entiende, de que reproduzcan las generalizaciones que comporte–, lo cual no ocurrirá si la reducción se articula mediante disyunciones abiertas, y que conviene también tener en cuenta la distinción entre las propiedades de instanciación de los estados mentales –aquéllas que poseen en virtud de peculiaridades, digamos, inocuas de su encarnación en un sistema en particular– y sus propiedades de diseño –aquéllas cuya instanciación viene sujeta a ciertas restricciones físicas y evolutivas, y que, *pace* Kim, pueden resultar proyectables aun cuando la disyunción de las propiedades físicas de sus misceláneas materializaciones no lo sea; acaso también que la proyectabilidad deba entenderse como una cuestión de grados –tal como se ha venido sugiriendo respecto del carácter

necesario o contigente de la identidad entre tal o cual clase de fenómenos y el conjunto de sus materializaciones físicas. El trabajo de Jacob, entretanto, nos enseñará otro modo de precipitar las veleidades eliminacionistas de Kim.

Antes de abandonar tan dilatado ejercicio sobre el jade y el dolor, habrá tiempo de tantear una última senda en la defensa del compromiso antirreduccionista del funcionalismo: la que, de la mano de esa reescritura de la noción de proyectabilidad en términos de grado que se ha venido esbozando, nos llevaría a aceptar que el concepto de jade pueda, a fin de cuenta, constituir una clase propia de alguna ciencia especial, y participar –acaso más humildemente que el de dolor en el ámbito de la psicología– en la formulación de generalizaciones válidas de esa disciplina distintas de las que pudieran construirse empleando la disyunción de jadeíta y nefrita: quedaría desarbolado, así, el ímpetu primero del vigoroso embate de Kim. También, tras tomar aliento, se escrutarán cuidadosamente las conclusiones que en cuanto atañe a la necesidad de fundamentar la autonomía de la explicación psicológica en una noción suficientemente potente de eficacia causal de lo mental puedan haber fructificado en esta inesperada controversia acerca del dolor y el jade.

De la mano de recientes, fulgurantes avances en nuestro conocimiento de la fisiología nerviosa de diversas variedades de actividad mental, la disputa sobre la naturaleza de los lazos epistemológicos que hayan de unir a la psicología con las ciencias del cerebro se ha prolongado bajo la forma de unas **Prácticas de taxonomía neurológica y psicológica** tras las que queda a veces velado un hondo replanteamiento de las nociones de **estructura y función**. Así, por ejemplo, un influyente trabajo de William Bechtel y Jennifer Mundale intentaría, ya en 1999, denunciar tres errores graves errores que habrían viciado el consenso antirreduccionista fraguado al amparo de Putnam: un uso espurio, desligado de la práctica científica, de los conceptos de estado mental y estado cerebral, la aplicación de un doble rasero en nuestras exigencias de precisión hacia los criterios de taxonomización empleados en el seno de la neurofisiología y los atribuidos a la teoría psicológica, y, por fin, una injustificada dicotomización de los niveles explicativos admisibles en la labor científica, anquilosada en torno a las nociones de estructura y función. Si hay, como someramente veremos, verdad pero escaso fruto en la primera acusación, mientras que se da por hecha, en la segunda, la falsedad de las tesis discutidas, habrá que detenerse en indagar como el esfuerzo por permeabilizar los límites entre estructura y función entronca con la labor que en el propio seno del funcionalismo venía desarrollando al menos desde 1987 William Lycan, y hacia la que ya en 1969 dejara anotado su temprano interés Herbert Simon. Esa relectura de las nociones de estructura y función acaso nos permita rescatar cuanto pueda haber de valioso en los argumentos contra el funcionalismo de Lawrence Shapiro, al margen del infructuoso intento de hacer reposar la verdad del reduccionismo sobre una distinción entre diferencias físicas relevantes y diferencias físicas irrelevantes que no hace sino apelar veladamente a criterios funcionales, tal como se estudiará primero en un caso imaginario –la dilatada polémica sobre el cerebro de silicio que ya puede encontrarse en Pylyshyn o en Searle– y luego en la interpretación de uno

experimental –los fenómenos de reorganización intermodal del córtex auditivo del hurón descritos por el equipo de Laurie von Melchner, que Shapiro quiere entender no como instanciaciones diferentes de idénticas funciones psicológicas en un mismo organismo, sino como ejemplo de diferencias físicas irrelevantes.

También Carl Gillett concluirá que las conclusiones reduccionistas en que confluyen Shapiro o Bechtel y Mundale descansan a fin de cuentas sobre falacias más o menos sutiles de petición de principios. De acuerdo con el análisis desplegado por Gillett, el fondo de la cuestión es que mientras el renovado brío del reduccionismo que surge a partir de las reflexiones de Kim depende de una concepción plana de la relación de instanciación –es decir, de la exigencia de que los poderes causales de la propiedad instanciada sean un subconjunto de los poderes causales de la propiedad en la que aquella se instancia y, por tanto, ambas propiedades se prediquen del mismo ente particular–, el antirreduccionismo se había construido sobre una concepción dimensional de dicha relación, en la que los poderes causales de las propiedades instanciadas pueden diferir de los que corresponden a las propiedades en las que aquellas se instancian precisamente en la medida en que unos y otros pueden pertenecer a diferentes entes –por ejemplo, a un organismo y a algunas de sus partes constituyentes. Se diría, entonces, que en esta distinción entre las propiedades físicas del individuo que alberga una determinada propiedad funcional y las de sus partes constituyentes acaso pueda hallarse tierra fértil para la irreductibilidad de la explicación psicológica. Entre *petitio* –esta vez con un marcado acento *ad autorictatem*– e *ignoratio elenchi* basculan los argumentos de John Bickle en este ámbito: cuando no se santifica la falsedad del antirreduccionismo apelando a que no es compatible con la práctica científica inspirada por tesis reduccionistas, se reconstruyen dichas tesis como principios heurísticos que, como tales, resultan tangenciales a la controversia, o se formulan en términos que hasta el más aguerrido antirreduccionista podría compartir. Más recientemente, Mark B. Couch ha intentado poner algún orden en el acúmulo de argumentos contra la interpretación antirreduccionista de la tesis de realizabilidad múltiple distinguiendo entre aquellos que tratan de mostrar que las propiedades funcionales de varios sistemas u organismos no son idénticas –contra las pretensiones del antirreduccionista– y aquellos que apuntan a que sí que son idénticas, después de todo, las propiedades físicas en las que dichas propiedades funcionales se materializan en los diversos sistemas u organismos, siempre que se clasifiquen con criterios neurofisiológicos adecuados. Anotaremos, tan sólo, que la prosperidad de la primera familia de argumentos, de no ir acompañada por igual suerte para la segunda, no vendría, una vez más, a alentar conclusiones reduccionistas, sino eliminacionistas.

Si, como venimos viendo, el ímpetu reduccionista debe vérselas con la ardua tarea de meter en vereda la contumaz disparidad de encarnaciones que pareciera poder tomar lo mental, una defensa de la autonomía de la psicología que aspire a desbordar el ámbito de lo pragmático no puede esquivar la obligación de rendir cuenta de alguna suerte de eficacia causal que sea propia de los estados mentales, y que lo sea en un sentido suficientemente sólido como para servir de sustento a un

conjunto de generalizaciones susceptibles de ser expresadas mediante el lenguaje teórico de la psicología, en virtud de determinadas peculiaridades que éste presente, pero no mediante el lenguaje teórico de la neurofisiología u otras ciencias más básicas. Entre los intentos más influyentes de encajar esa noción de causalidad mental en nuestra concepción global de la causalidad cobra particular relieve el que desde 1970 viene articulando Donald Davidson, quien ha señalado infatigablemente que nuestras prácticas taxonómicas en lo que atañe a estados mentales gravitan en torno a criterios normativos, canónicos, de índole racional –como el principio de caridad–, de suerte que, en tanto venga descrito y clasificado en términos psicológicos, cualquier estado mental se verá envuelto en regularidades nómicas que es imposible recoger en un vocabulario teórico que, como el de la física o la fisiología, permanezca ajeno a dichos cánones. Aunque los estados mentales traban las relaciones causales que traban –piensa Davidson– en tanto que no son sino estados físicos, las regularidades que emergen bajo su clasificación psicológica no pueden expresarse mediante leyes físicas –sólo, dicho sea de paso, bajo leyes psicológicas *caeteris paribus*–, como no pueden quedar reducidas a propiedades físicas las propiedades psicológicas que quedan delineadas bajo tales criterios taxonómicos: lo mental, en suma, es físico, pero su naturaleza física es anómala. De lo que se trata, pues, es de dilucidar si las *razones* a las que solemos apelar en la explicación psicólogica –razones para actuar, para pensar o desear…– son, a fin de cuentas, genuinas *causas*: si, como las **Obras**, son amores **o** sólo **buenas razones**; si, dicho de otro modo, el principio de **caridad** es un antídoto eficaz **contra** el principio de **herencia causal** sobre el cual descansa el embate fisicalista.

No sólo el principio de herencia esgrimido por Kim quedaría en cuestión: la propuesta de Davidson nos obligaría a revisar también una veta de la concepción materialista del mundo tan nuclear como es la idea de exclusión causal –que ningún fenómeno físico puede tener más de una causa física suficiente ni, por tanto, más de una explicación física completa e independiente–, que no es difícil contraponer a la constatación de que, en el ámbito de la vida mental, múltiples explicaciones de los mismos hechos acostumbran a convivir, no siempre plácidamente. Al igual que ocurre al hilo de la distinción entre causas estructurantes y causas desencadenantes introducida por Fred Dretske con el propósito, también, de encontrar cobijo para la eficacia causal de lo mental, veremos que la maniobra de Davidson liga estrechísimamente la posibilidad de hallar dicho cobijo a la noción de intencionalidad. Pero, sobre todo, conviene advertir que Davidson nos está pidiendo, como ya hiciera David Hume, que diluyamos la noción de relación causal en la de regularidad nómica. Así, que un suceso sea la causa de otro habrá de consistir en el hecho de que el primero pertenezca a un tipo tal que sus miembros regularmente anteceden, de acuerdo con un enunciado legaliforme, a los del tipo al que pertenece el segundo; será preciso, claro, fijar las condiciones que distingan a esos enunciados con rango de ley de meras expresiones de regularidades cuya categorización como reflejo de relaciones causales vendría a ser tanto como una *reductio* de la posición de

Davidson, pero el paso crucial se ha dado ya en la medida en que la epistemología de las leyes se ha convertido en la ontología de las relaciones causales, y viceversa.

De hecho, el principio de herencia causal sobre el que descansa la reivindicación del reduccionismo auspiciada por Kim es en su origen un intento de desarbolar la concepción de la causalidad –*ergo*, de lo mental– que se acrisola en este monismo anómalo; Davidson, por su parte, acusará a Kim de ignorar las consecuencias de la distinción entre casos y tipos al no reparar en que las relaciones causales que establezca un estado mental del tipo *M* encarnado en un estado nervioso del tipo *P* no tienen por qué restringirse a las que pertenezcan a dicho estado nervioso por mucho que ambos sean un único y el mismo fenómeno. Puesto que el estado mental y el estado nervioso son idénticos, sabemos que tienen –o mejor, tiene– el mismo conjunto total de propiedades, pero no se sigue de ello que el subconjunto de esas propiedades en virtud de las cuales hemos asignado al tipo psicológico *M* sean las mismas que aquellas por las que lo hemos asignado al tipo neurofisiológico *P*. Lo esperable, antes bien, es que la clasificación como *M* atienda a propiedades psicológicas, es decir, tejidas en una urdimbre de racionalidad –y, con ello, de normatividad, intencionalidad, acaso subjetividad…–, mientras que la clasificación de ese mismo estado como *P* repose sobre propiedades neurológicas ajenas a dicho tejido, de modo que cada práctica taxonómica haría brotar su propio horizonte de generalidades nómicas –o sea, de relaciones causales–, invisible desde la perspectiva de la otra. En un examen cuidadoso de la noción de superveniencia estricta perfilada por Kim veremos cómo cierta propensión a pasar por alto la distinción entre casos y tipos, semejante a la denunciada por Davidson, y con ello a deslizarse hacia la petición de principios contra las tesis antirreduccionistas; asimismo, al margen de la mayor o menor robustez de las conclusiones de ese análisis, será preciso subrayar que si es la identidad entre lo mental y lo físico lo que pretendidamente hace que lo mental haya de ceder su eficacia causal en tanto que mental, también lo físico habrá de cederla en virtud de su identidad con lo mental, así que nos veremos forzados a elegir entre la conclusión de que ni lo físico ni lo mental atesoran eficacia causal alguna y la de que ambos lo hacen, por así decir, a su manera.

Pese a que veremos luego a Davidson ensayar una débil defensa de la eficacia causal de lo mental en el terreno de la concepción de la causalidad delimitada por Kim, es en el intento de ahondar en las diferentes concepciones de la causalidad que subyacen a los argumentos de uno y otro, y de sopesar la coherencia de cada una de ellas con las prácticas de atribución de causalidad habituales en el trabajo científico, en lo que habremos de demorarnos antes de concluir que la reconciliación del carácter relacional y el carácter causal de lo mental anhelada por el funcionalismo sólo será completa si se acompaña de una profunda reelaboración de nuestra noción de causalidad, que todo indica que dicha reelaboración habrá de consistir en una desustancialización de las causas tan severa como aquella a la que, desde los tiempos del conductismo lógico, venimos sometiendo a la mente y que, en cualquier caso, no parece que la cuestión de la naturaleza de lo mental y la de la naturaleza de

causalidad vayan a aclararse del todo mientras no lo hagan a la par. Como poco, si hemos de preservar una noción de eficacia causal de lo mental suficientemente robusta como para acoger en su seno una psicología autónoma, refractaria a la reducción, pero no estamos por la labor de admitir bajo nuestra noción de causalidad a propiedades funcionalmente definidas como no sea en virtud de su identidad con propiedades físicas, no nos quedará más remedio que aplicar la idea de realizabilidad múltiple a la propia noción de causalidad, que viene a ser tanto como dar por buena la tesis de que distintas reconstrucciones de la relación entre causas y efectos permiten, a través de distintos vocabularios teóricos, alumbrar distintas generalizaciones sobre el mundo que nos rodea.

Distintos vocabularios teóricos son, pues, como distintos **Aparejos para apresar lo mental**, y distintos aparejos ofrecen distintas cobranzas. La defensa de la autonomía de la psicología –es decir, de la diversidad de aparejos que es imprescindible en esa tarea– es una de las señas distintivas del cognitivismo, como lo es de la concepción funcionalista de lo mental que acompañara su nacimiento. Si ya Skinner entendía cabalmente el conductismo como una filosofía de la ciencia –incluso como una gnoseología–, el cognitivismo se perfila como una filosofía de la ciencia que aspiraría a dotar a la psicología de un mayor grado de autonomía respecto de otros saberes del que pudieran ofrecerle el conductismo o el reduccionismo neurofisiológico al uso. Pero frente a la renuencia o la insconstancia con que Skinner afrontara los compromisos ontológicos que se derivaban de sus planteamientos, el cognitivismo –que además de con Skinner se fragua en el diálogo con Ryle– trata de hacerse valer como una toma de posición francamente ontológica, si bien en un sentido muy peculiar: como una concepción de lo mental ontológicamente neutral a la luz de la cual la autonomía de la psicología quedaría purgada de vestigios dualistas. Ahora bien, esta porfiada defensa de una psicología autónoma ha podido ser confundida con un malaconsejado desprecio de la investigación sobre la estructura y función cerebral, que podía barruntarse en ciertas proclamas de voces tan influyentes como la de David C. Marr –pese a que, como se apuntará, una actitud así se hallaba patentemente alejada tanto de su trabajo científico como de sus reflexiones, y pese a que su defensa de la autonomía de la psicología era en todo caso más bien titubeante–, o la del propio Putnam. Sea como sea, la confusión entre autonomía de la psicología e irrelevancia de la neurofisiología ha podido avivar los ánimos –qué duda cabe– contra toda lectura antirreduccionista del funcionalismo. De la mano de Harty H. Field, por tanto, no estará de más que procuremos trazar nítidamente las fronteras que separan una tesis de otra; la discusión de un argumento de Gary Hatfield nos ayudará a ver que hacerlo con propiedad exige indagar a fondo en la noción de equivalencia funcional y los distintos alcances que cabe darle, y eso alumbrará el peligro que supone para el cognitivismo adoptar un criterio de equivalencia funcional excesivamente rígido, que lo abocaría bien a dar alas, después de todo, a una interpretación reduccionista del papel de la explicación psicológica, bien al desmoronamiento hacia lo idiográfico que ya hemos vislumbrado en el funcionalismo analítico.

Pero de manera mucho más rotunda que en Marr, donde la defensa de la autonomía de psicología cobra cuerpo, y donde madura la idea de que distintos vocabularios teóricos han de proporcionar distintas cosechas explicativas, es en el trabajo de Pylyshyn, a cuyos ojos la insistencia en una visión reduccionista de las relaciones entre la psicología y las ciencias del cerebro, una vez que hemos aprendido que estados mentales del mismo tipo pueden diferir en su encarnación física, es contraria al mandato de capturar tantas generalizaciones como nos sea posible que define la labor del científico –como también lo sería, veremos, despreciar las cuantiosas mieses que la neurofisiología puede ofrecer. Apartándose abruptamente de la forma en que el positivismo había concebido la relación entre *explanandum* y *explanans*, en un giro sin duda inspirado por Fodor, Pylyshsyn viene así a deslindar unas disciplinas de otras no en virtud de su objeto de estudio o el contenido de sus enunciados, sino de ciertas propiedades conceptuales del vocabulario que emplean. Pero recurrir a un determinado vocabulario teórico, con sus propios recursos conceptuales, trae consigo una cierta taxonomía de los fenómenos que conforman el *explanandum* de la teoría –a la vez que los delimita–, y dicha taxonomía hace posible la detección y explotación de generalizaciones que otros vocabularios teóricos ocultarían –a la vez, claro, que vela otras que serían visibles bajo taxonomías propiciadas por vocabularios diferentes. Que diferentes vocabularios teóricos puedan alumbrar diferentes generalizaciones se destila en una perspicaz observación de Pylysyhn sobre la lógica de la descripción y la explicación: los enunciados sobre explicaciones generan contextos referencialmente opacos, de suerte que una descripción de un fenómeno en un vocabulario teórico puede constituir una explicación de ese hecho mientras otras descripción del mismo hecho, igualmente verídica pero expresada en otro vocabulario teórico, no alcanza a explicarlo –puesto que explicarlo es, precisamente, entretejerlo en una urdimbre de generalizaciones. Pero lo que sustenta el don de atrapar generalizaciones vedadas al vocabulario de la neurofisiología que atesora la explicación psicológica es –piensa Pylyshyn– el sencillo hecho de que ésta nos permite describir estímulos, estados internos o conductas tal como son *interpretados* por el propio sujeto, al margen de sus propiedades objetivas, e incorporarlos a taxonomías basadas en ese marco subjetivo. Sería, en suma, la carga semántica de los conceptos teóricos articulados por la psicología, su –por así decir– hondura intencional, lo que haría irremplazable la cosecha que ofrecen.

Desde un punto de vista levemente distinto, y sólo levemente esbozado por Pylyshyn, podemos ver la diferencia entre descripción y explicación como el resultado de aplicar a las propias relaciones causales la distinción entre casos y tipos que anida en las raíces del funcionalismo: la descripción de un caso concreto de relación entre un hecho que describimos como causa y otro que describimos como efecto, aun siendo verídica, no alcanzaría a explicarlo, entonces, mientras no lograse imbricarlo en el tipo de relaciones que se dan entre un tipo de causas y un tipo de efectos. Dicho de otro modo, también de aire ya familiar: que toda relación causal concreta sea un fenómeno físico no entraña que toda clase de relaciones causales lo sea, en el sentido de que pueda delimitarse empleando el lenguaje de la física. De

nuevo, vistas así las cosas, es una cierta reelaboración de nuestra idea de causalidad, capaz de acomodar una noción de eficacia causal de lo mental suficientemente robusta, lo que subyace a la reivindicación de la autonomía de la psicología.

No nos será difícil reconocer en el pensamiento de Pylyshyn las huellas de las críticas que Chomsky opusiera, ya en 1959, al programa de investigación sobre la conducta verbal bosquejado dos años antes por Skinner –ya, de hecho, las habremos rastreado en el las propuestas de Miller, Galanter y Pribram. Tampoco es particularmente arduo escuchar el eco de Chomsky en la controversia en que Pylyshyn se enzarza con el realismo perceptivo directo defendido por James J. Gibson (1950, 1966, 1979). Sin embargo, es de rigor señalar que la caracterización de estímulos y respuestas que Skinner preconizaba no se amolda bien al esquema de descripción fisicalista, cinemática, al que Pylyshyn asemeja el vocabulario teórico del conductismo: bajo el recurso subrepticio a las proscritas nociones mentalistas del que Chomsky –con toda razón– acusara a Skinner se esconde una decidida propuesta de definición funcional de estímulos y respuestas –más explícitamente: de clasificación de estímulos y respuestas en tipos atendiendo a criterios funcionales– que cabe reconocer, seguramente a su pesar, como una veta del pensamiento de Skinner en la que se anticipan algunos rasgos cruciales de la concepción cognitivista de lo mental, al igual que sucede –decíamos– con su certera comprensión de que el conductismo es ante todo una filosofía de la ciencia.

Si en el libro VI de su *Metafísica* dejó escrito Aristóteles que son muchas las maneras en que puede entenderse el ser, el libro I de la *Ética* que dedicara a su hijo Nicómaco se abre con la constatación de que son también muchas las maneras en que puede entenderse el propio acto de entender, o de explicar. De aquellas palabras ha venido manando desde entonces –pero sobre todo en el último siglo, como en una pugna ante la pujanza del positivismo– una caudalosa reflexión sobre los distintos modos en que cabría decir que algo queda explicado. Con el tiempo, esos torrentes han acabado sedimentándose en un debate **Sobre** las nociones de **explicar y comprender** cuyo eco en los intentos de sufragar la autonomía de la psicología articulados por Davidson y Pylyshyn es tan claro que no cabe dejar de señalarlo siquiera a vuelapluma. La distinción, en efecto, entre *Erklären* y *Verstehen* –esto es, entre explicación y comprensión– en la que ya en 1858 tratara Johann Gustav Droysen de enraizar un meticuloso análisis de las diferencias entre *Naturwissenchaften* y *Geisteswissenschaften*, tamizada después con las aportaciones de Wilhelm Dilthey, Max Weber o Robin G. Collingwood, es la plantilla sobre la que se dibujaría, de la mano de Georg Henrik von Wright, una reivindicación de ciertos modos artistotélicos de explicación, de corte finalista e irreductiblemente teleológico, repudiados por el positivismo frente a las explicaciones mecánicas de la ciencia galileana. Aun cuando en las líneas de defensa de una concepción funcionalista de lo mental capaz de resguardar a la psicología del empuje reduccionista que parten del trabajo de Putnam y Fodor se haya tratado, por regla general, de rehuir cualquier filiación con la idea de explicación teleológica –como se ha tratado también de esquivar el matiz subjetivista que el concepto de comprensión cobrara desde

Dilthey–, es innegable que los argumentos de von Wright acogen en su seno buena parte de las ideas cruciales que Putnam y Fodor habían comenzado ya a entretejer. Sin ir más lejos, el decisivo papel de la distinción entre afirmaciones sobre estados mentales con alcance de casos y con alcance de tipos, que hemos visto ya prefigurada en ciertos aspectos del trabajo de Skinner, está también claramente presente en el aparato formal que sirve de base a la propuesta de von Wright, cuando éste nos advierte de que sólo estados de cosas lógicamente independientes entre sí, y tomados *en sentido genérico*, son susceptibles de quedar enlazados por relaciones causales. En parecidos términos –veremos– la distinción entre un concepto de descripción de naturaleza extensional y uno de explicación, teñido de intensionalidad, que es clave en el pensamiento de Pylysyhyn, subyace, muy velada, a la constatación de von Wright de que la validez del silogismo práctico, entendido como esquema básico de explicación psicológica, está supeditada a una cierta descripción de las conductas en él mencionadas; como en Pylyshyn –o, claro está, como en Davidson– esa constatación remite a la de que la explicación psicológica reposa sobre descripciones o clasificaciones intensamente impregnadas de criterios poco afines al vocabulario fisicalista –teleológicos, intencionales, racionales. Es, en suma, la naturaleza de los conceptos empleados en nuestra descripción o nuestra taxonomía de los fenómenos lo que determina que diferentes explicaciones permitan rendir cuenta de diferentes hechos acerca de tales fenómenos o capturar diferentes generalizaciones. Además, en la estela de Peter Winch o Kenneth Pike, von Wright no vacila en subrayar que entre los rasgos de la acción humana que más la alejan de los esquemas positivistas de explicación científica se cuenta el señorío que sobre ella ejercen las convenciones y normas sociales –una observación cuyas huellas en Davidson o Pylyshyn no son, por atenuadas, menos nítidas.

Los lazos entre el litigio de von Wright con el positivismo y la égida de la explicación psicológica emprendida por Putnam y Fodor pueden rastrearse hasta una consideración previa a todo escrutinio de la peculiaridad de la acción humana: von Wright apunta, en efecto, que el mero acto de subsumir un fenómeno bajo una generalización legaliforme que lo liga a otro no mitiga del todo nuestra necesidad de explicarlo, que habrá de calmarse bien mediante una investigación empírica sobre las causas eficientes de la relación entre los fenómenos cuya concomitancia describe la ley en cuestión, bien mediante la concesión a dicha relación del rango de elemento definitorio de dichos fenómenos, admitiéndola como integrante del significado de los términos que designan a la clase natural a la que los asignamos. Ésas, precisamente, han venido siendo las dos rutas cardinales de desarrollo de la concepción funcionalista de lo mental: la que, en el seno del funcionalismo analítico, pasa por buscar el sustento de los patrones de relaciones funcionales que se proponen como constitutivos de los estados mentales en el análisis del significado del vocabulario mentalista y la que, de mano del funcionalismo empírico, para por hacerlo entre los hallazgos cosechados por la psicología científica. Ahora bien, no sólo más acá de la peculiaridad de la acción humana, sino también más allá, encontraremos que en las reflexiones de von Wright y las de Davidson o Pylyshyn se entreoye parecido

murmullo: la conclusión acaso de mayor alcance a la que conduce la investigación de von Wright es que hemos de reconstruir nuestra noción de causalidad para hacer transparentes ciertas vetas que la recorren, en las que se hallan cristalizadas las ideas de posibilidad, necesidad y acción intencional –es decir que, a su juicio, tal vez sea finalmente la noción de causa mecánica la que descanse sobre la de causa teleológica–; también la defensa de la autonomía blandida por Davidson y Pylyshyn acababa por alentar, como se ha dicho, una reelaboración de nuestra concepción de la causalidad cuyos perfiles, desde luego, difieren en cada caso, pero también muestran evidentes coincidencias. Con todo esto, se hará ineludible recordar el temprano alegato en el que, ya en 1892, William James comenzara a hablar de una psicología científica madura como aquella por cuyos goznes se filtrarían constantemente las aguas de la crítica metafísica –todo lo contrario de una disciplina, como por lo común se ha pretendido desde entonces, libre de excrecencias filosóficas.

Mientras cifraba en la caracterización funcional de los estados y procesos psicológicos la esperanza de una cabal comprensión de la naturaleza de lo mental, el funcionalismo ha venido mostrándose sumamente reacio a extender la misma minuta a la descripción de los estímulos y las respuestas con los que aquellos se ligan inextricablemente: así, incluso en los ejemplos paradigmáticos de caracterización funcional de autómatas sencillos se opta de manera decidida por una descripción categórica, generalmente en términos intuitivos o vagamente físicos, de las aferencias y las eferencias del sistema del que se trate. El empeño en esquivar toda caracterización de estímulos y respuestas que no fuera la estrictamente física era patente ya en los trabajos con que Armstrong diera cuerpo a la concepción fisicalista de la mente que se había gestado en las reflexiones de Place sobre los límites del análisis disposicional; en ese sentido, pues, el funcionalismo no se habría apartado sustancialmente de la teoría de la identidad psicofísica. Cuando Block desecha la posibilidad de que estímulos y respuestas deban ser descritos en términos funcionales –es decir, del mismo modo que se propone abordar la caracterización de los estados mentales– comparte además con Armstrong uno de los motivos que alientan su rechazo: el temor de que dilatar el recurso a la definición funcional abarcando en su seno también **Los lazos con el mundo** nos aboque a una irremediable circularidad, a fórmulas desprovistas de anclaje fuera de sí mismas, libres para girar en el vacío. Sin embargo, la cuestión de **cómo describir estímulos y respuestas** encierra a ojos de Block secuelas más devastadoras para el propio funcionalismo, ya que deja ver que, igual que el fisicalismo o el conductismo, éste es incapaz de construir herramientas que permitan dilucidar en qué circunstancias es procedente la atribución de estados mentales a otros organismos o sistemas sin incurrir en los vicios contrapuestos del liberalismo –hacerlo con desmedida prodigalidad– y el chauvinismo –negar la posesión de estados mentales a otros por la razón espuria de que no sean suficientemente parecidos a nosotros–, a los que están condenados, a su juicio, conductismo y fisicalismo respectivamente.

Veremos, no obstante, que las dificultades en que el funcionalismo pueda verse a la hora de dirimir cuándo la atribución de vida mental está justificada no se

ven afectadas en lo más mínimo por la decisión de emplear o no un vocabulario funcional para la descripción de estímulos y respuestas, sino que se dan ya en toda la envergadura que puedan alcanzar incluso si se respeta escrupulosamente la restricción de tal vocabulario al ámbito de los estados internos tal como es propugnada por Block. Pero además de inútil, mantener dicha restricción es gravoso: el intento de expulsar las referencias a otros estímulos, estados internos o respuestas de la delimitación de cada tipo de estímulo o respuesta que podamos incorporar a una teoría funcionalista conduce a una dinámica que habrá ya de resultarnos familiar, en la que las referencias desterradas reaparecen aquí y allá de manera más o menos encubierta igual que reaparecían disfrazadas, precisamente en el vocabulario empleado para describir estímulos y respuestas, las alusiones mentalistas que Skinner había vetado de su explicación de los procesos de aprendizaje. La obstinación en mantener la caracterización de estímulos y respuestas al margen del expediente de definición funcional aboca al funcionalismo, en suma, a enfrentarse, en esencia, a la misma acusación que Chomsky lanzara sobre Skinner.

Si claro es el vínculo entre el funcionalismo y el fisicalismo en cuanto al vocabulario elegido para la descripción de estímulos y respuestas, no menos palpables son las raíces que hunde en el seno del conductismo el rechazo funcionalista a que dicha descripción pueda teñirse de matices subjetivos –rechazo en el que, pese a todo, anida la propia idea de definición funcional. Es frecuente, en efecto, perfilar el germen del proyecto de Watson como su ruptura con la tradición, que se remontaba hasta Conwy Lloyd Morgan y su *doble inferencia*, de añadir al informe objetivo sobre estímulos, asociaciones y respuestas una descripción subjetiva de la experiencia, de lo que de aquellos estímulos, asociaciones y respuestas trasluce en la consciencia. La advertencia de que los conductistas tendían sin embargo a impregnar de esos matices mentalistas el vocabulario que pretendían objetivo se había comenzado a oír ya en 1954 –de voces tan autorizadas como las de William K. Estes, Sigmund Koch, Kenneth MacCorquodale, Paul E. Meehl, o Willam S. Verplanck–, y la convicción de que estímulos y respuestas habían de verse sometidos a una definición en términos funcionales, según sus relaciones recíprocas, acompaña a Skinner desde al menos 1931. Es en virtud de esa táctica, de hecho, que Skinner confiaba en redimir para la ciencia la ley del efecto formulada por Thorndike, depurándola del subjetivismo residual que Watson había denunciado y construyendo sobre ella la noción de condicionamiento operante. Comoquiera, en cualquier caso, que Skinner estaba tan convencido de admitir en el vocabulario del análisis de la conducta a definiciones funcionales de eventos externos al organismo como de repudiar sin miramientos cualquier referencia a sus estados internos, la descripción de estímulos y respuestas acabaría tornándose tan abiertamente funcional como veladamente mentalista.

Acaso haya de encontrarse el origen del profundo rechazo que el funcionalismo parece guardar hacia la definición funcional de estímulos y respuestas, al menos en parte, en el carácter de mito fundacional que revisten las objeciones de Chomsky al modelo de explicación fraguado por Skinner y, por tanto,

en una suerte de antagonismo residual, ya que dichas objeciones descansan sobre el veto de Skinner a un vocabulario mentalista cuyo retorno no lograría evitar, pero son independientes de su recurso a la definición funcional de estímulos y respuestas. Sea como sea, la controversia entre Chomsky y Skinner impulsó sin duda la reflexión sobre cuál fuese el vocabulario idóneo para la caracterización de estímulos y respuestas: sólo un año después, en la idea de bucle de retroacción articulada por Miller, Galanter y Pribram, había de formarse, como veremos, la semilla en la que germinaría el procedimiento de definición funcional simultánea –la idea que serviría a Putnam para conectar los avances en teoría de autómatas con la pregunta por la naturaleza de lo mental.

La definición funcional de estímulos y respuestas, en suma, forma tanto parte de los orígenes de la psicología cognitiva y de la concepción funcionalista de lo mental como lo hace su rechazo. De hecho, no es difícil adivinar, por ejemplo, en la incansable defensa de la autonomía explicativa de la psicología en la que Fodor no ha cejado, cómo el carácter velada cuando no abiertamente funcional de las nociones de estímulo y respuesta relevantes para la teorización psicológica se convierte en un argumento más contra la reducción de la psicología a ciencias más básicas. Provista de esa poderosa herramienta, entonces, conjurados si pudieran conjurarse los peligros de circularidad y de irrefrenable prodigalidad en la atribución de vida mental que preocupaban a Block, la psicología se vería en disposición de afrontar la definición de estímulos y respuestas con horizontes más amplios de los que le proporciona el funcionalismo analítico –ceñirse a la descripción ingenua de estímulos y respuestas que forme parte del sentido común, la cual, por otro lado, no sería raro que resultara, después de todo, ser de naturaleza funcional– o el funcionalismo empírico –confiarlo todo a su descripción física o, a lo sumo, fisiológica, si es que fuera posible desplazar la frontera que intuitivamente separa a la mente del mundo para dejar fuera, junto a lo que venimos entendiendo por estímulos y respuestas, a las señales sensoriales y motoras.

La cuestión de cómo debemos describir estímulos y respuestas ofrece, así pues, un puerto propicio para aventurarse en algunas de las diversas problemáticas que en torno a la concepción funcionalista de la mente atañen a la difícil tarea de escrutar cómo se incardina **El mundo en la mente y viceversa** –de trazar las lindes precisas, diríamos, entre la mente y el mundo del que forma parte. Aceptar la propuesta del funcionalismo empírico de que el vocabulario que empleemos para caracterizar estímulos y respuestas, a fin de soslayar los peligros que comporta su descripción funcional, asuma como referente la actividad nerviosa de transductores y efectores –es decir, estímulos y respuestas entendidos en sentido proximal– es tanto como comprometernos con una concepción internista de la vida mental –y, por ende, de la explicación psicológica–, en la que sólo alcanzar a intervenir las propiedades de nuestras representaciones, nunca las de las cosas del mundo a las que éstas se refieran.

Habiendo transitado en sentido contrario el mismo trayecto –los compromisos internistas del funcionalismo y el cognitivismo tempranos implican un compromiso

no menos rígido con la caracterización de estímulos y respuestas en términos proximales–, supo ver Dennett que recorrerlo entraña tratar a los órganos de los sentidos como oráculos cuyos designios se adentran en lo irremediablemente desconocido, y que en esa extraña imagen palpita ya la idea del cerebro como una maquinaria enteramente ajena a la semántica de sus propios estados, que pendería inerte de unos procesos regidos sólo por su inexorable sintaxis interna. Pero sería Block, poco después, quien con mayor detenimiento escrutaría las consecuencias que guardaba para nuestra comprensión de los lazos entre psicología y semántica distinguir entre el contenido de una creencia o un deseo entendido en sentido amplio –esto es, como dicta el sentido común, en virtud de sus relaciones con la estimulación distal– y su contenido en sentido restringido –delimitado únicamente a tenor de sus relaciones con la estimulación proximal. Si del contenido restringido de los estados mentales cabía pensar que mantuviese una dócil superveniencia con el estado físico del organismo o sistema que los albergara, difícil sería esperar lo mismo del contenido amplio, que, como se hizo común decir, *no está en la cabeza*; si el contenido amplio enlazaba con la noción semántica de referencia, o *Bedeutung* –y, de su mano y la de Frege, con la de valor de verdad–, el contenido restringido parecía evocar la noción de sentido, o *Sinn*; si el papel del contenido amplio en la explicación psicológica se veía limitado a los casos en los que careciésemos de información acerca de los estados mentales del sujeto, el contenido restringido se erigía como la verdadera clave de arco de la psicología. Entre experimentos imaginarios sobre *Doppelgänger* y mundos gemelos, poco a poco iría quedando bosquejado lo que la psicología requeriría de una teoría semántica: fundamentalmente, que fuera capaz de rendir cuenta de la distinción entre significado amplio y significado restringido, del papel de cada uno en la explicación de la conducta, y de los mecanismos por medio de los cuales el significado restringido podía, dado un contexto, generar un significado amplio y un valor de verdad.

Que el contenido de una actitud proposicional, entendido en sentido restringido, verse acerca del entorno del organismo o de propiedades fenomenológicas de los objetos de dicho entorno, son posibilidades que ha tratado de despejar McDermott, con el ánimo de convencernos de que no hay otro horizonte para la incorporación de estímulos y respuestas a la teoría que su descripción proximal. Si bien todos negaríamos albergar, por ejemplo, creencia alguna la radiación electromagnética que incide sobre las células de nuestras retinas –lo que quiera que creamos, se diría, es acerca del color de las cosas–, esta opacidad es connatural a las actitudes proposicionales *de re*: negamos creer tal o cual cosa acerca de la longitud de las ondas electromagnéticas que alcanzan nuestra retina, aunque no acerca del color de esto o aquello, por los mismos motivos que Edipo niega que desee casarse con su madre, aunque no que desee casarse con Yocasta. Aunque esta concepción de creencias y deseos se aparta de la que dicta el sentido común –concede McDermott–, no lo hace hasta el punto de quebrar su compromiso con el realismo. Sin embargo, los escenarios modales en los que descansa la argumentación de McDermott –nuevas andanzas por la Tierra Gemela– la hacen vulnerable a una

enmienda general: si consideramos que la verdad de una creencia acerca de un objeto no depende de dicho objeto sino más bien de dicho objeto en el mundo actual o de su trasunto epistémico en otros mundos posibles, no nos veremos forzados a concluir que dependa del heraldo proximal que comparece ante los órganos de nuestros sentidos. Tendríamos entonces una teoría psicológica capaz de reconstruir nuestras creencias y deseos como actitudes proposicionales con contenido restringido acerca del entorno del organismo, semejantes en todo a las creencias y deseos del sentido común –clasificadas según su contenido amplio– excepto en su relación con la verdad en entornos modales. Pero una psicología así sería para McDermott una psicología estéril, cuyas generalizaciones habrían quedado quebradas por la torsión entre un término concerniente a estímulos o conductas descritos en términos proximales y otro que atañería a características del entorno –una torsión que bloquearía, so pena de *equivocatio*, la extracción del consecuente. Pues bien: asumir un vocabulario funcional en la descripción de estímulos y respuestas abre una ruta que podría servirnos para escapar de la desalentadora conclusión de McDermott y que –veremos– parece de hecho no alejarse mucho del camino seguido por Pylyshyn en su ensayo de fundamentación de la teoría cognitiva –se trataría, en esencia, de restringir el veto sobre del vocabulario funcional al ámbito del procesamiento perceptivo temprano y al del control motor, liberando de esa traba al resto de la psicología. Tampoco –se argumentará– perderíamos así el rumbo que para evitar las escolleras cartografiadas por Block dejara trazado Bechtel, y que nos aboca a responder en términos teleológicos a la pregunta por la atribución de estados mentales; antes al contrario, incorporar a la caracterización de su vida mental determinadas consideraciones acerca de cómo el comercio con su entorno que mantiene un organismo le permite afrontar sus necesidades es ya *in nuce* emprender una caracterización funcional de estímulos y respuestas. De hecho, si se pretendiese dar a esa relectura teleológica del funcionalismo una interpretación homuncularista, a la manera de Dennett, nos encontraríamos con que aquella liberalidad que Block encontraba exorbitante se nos habría vuelto imperativa: no es descabellado pensar que sólo bajo una caracterización funcional de estímulos y respuestas pueda una intrincada red de células nerviosas convertirse en un homúnculo.

Contra esa escisión de mundo y mente que levanta entre ambos **Un cerco invisible** cuyos contornos venimos punteando ha trabajado con celo Tyler Burge, quien dejó moldeados varios escenarios en los que ciertas diferencias en los hechos mundanos conllevan diferencias en creencias o deseos sin que el estado del organismo registre ningún cambio. Pensar que eso entrañe la quiebra del materialismo es ignorar –insiste Burge– la diferencia entre las relaciones de composición o causación, por un lado, y las de individuación, por otro, que cristaliza, entre otras cosas, en que el carácter local de aquellas no corresponde a ésta. Lo que encontraremos al observar el trabajo de los psicólogos, si Burge está en lo cierto, es bien distinto de lo que el internismo –el solipsismo metodológico, o, como prefiere decir Burge, el individualismo– decreta: una y otra vez al sujeto de sus teorías se le atribuyen estados mentales que determinan su comportamiento en virtud de su

semántica, la cual viene fijada por hechos tan distantes de su piel como ancho es el mundo donde acaecen. Desprovisto, entonces, del respaldo que le diera tal parentesco con el materialismo, y visto que resulta de hecho inexacto como descripción de las prácticas de construcción de teorías psicológicas, el individualismo deja entrever su carácter arbitrario y revisionista. En último término, el proyecto individualista reposaría sobre una reformulación del concepto de conducta en virtud de la cual las relaciones del organismo con su entorno que no se ciñeran a la exigencia de haber cristalizado en la fisiología dejaran de formar parte del objeto de estudio de la psicología, entendida como ciencia de la conducta en ese sentido injustificadamente tasado.

Algunos de los baluartes principales que jalonan la defensa del modelo internista de explicación psicológica frente al asedio de Burge quedaron cimentados en la pronta respuesta de Pierre Jacob. La objeción de Jacob es, en síntesis, que los argumentos de Burge toman como propiedades de ciertos estados mentales lo que no son sino propiedades de las expresiones lingüísticas que empleamos para adscribirlos a otros, o a nosotros mismos, de suerte que al mudar al sujeto a una comunidad lingüística diferente dichos estados mentales parecen verse envueltos en una metamorfosis que en realidad no les atañe. Esa confusión, además, inhabilitaría a una psicología articulada sobre los planteamientos de Burge para dar cuenta del hecho de que abrigamos –por otra parte, tan a menudo– creencias erróneas acerca del mundo que nos rodea, en la medida en que serían los propios hechos del mundo los que operarían como constituyentes de nuestras creencias. Ciertamente, la mera constatación de la posibilidad de equivocarnos hiende una fisura entre el mundo y nuestra noción de él en la que el internismo tiene más fácil encontrar apoyo que el externismo –de hecho, como se apuntará, la idea de que albergar una creencia errónea sea en realidad creer erróneamente que albergamos cierta creencia no ofrece al externismo el bálsamo que a veces se ha querido ver en ella, pues conduce abruptamente a un *regressus*. El problema del error, como veremos, pesa en favor de esa concepción restringida de la conducta que Burge ha repudiado como artificiosa, ya que es precisamente bajo tales restricciones como cabría rendir cuentas de aquellas conductas cuya causa resultara ser un error –dicho de otro modo, si la presencia de los propios hechos del mundo como constituyentes de mis creencias me impidiera tener creencias falsas, también me impediría, *a fortiori*, actuar en consecuencia. Por último, es propósito de Jacob restaurar los límpidos nexos entre los estados mentales de un organismo y sus estados fisiológicos que las tesis de Burge, pese a su insistencia en desligar las cuestiones de composición y causación de las de individuación, habían perturbado.

No es sólo, sin embargo, el conocimiento sobre el mundo atesorado por una comunidad y destilado en la redoma del lenguaje lo que puede filtrarse hasta calar en la naturaleza de las creencias o los deseos de quienes forman parte de esa comunidad: también, junto al lenguaje, el mismo don parece poseer la historia propia de las cosas que deseamos, o en torno a las que orbitan nuestras creencias, y de nuestros vínculos con ellas. Lo que el lenguaje y la historia van imponiendo, en

definitiva, es la duda de que las representaciones internas sobre las que operan los procesos computacionales que postula el psicólogo puedan seguir identificándose con las creencias o los deseos que saturan nuestro día a día o con el sustrato de su semántica. Esa divergencia entre al aparato conceptual de la psicología cognitiva y el de la psicología coloquial ha sido explorada por Lynn Rudder Baker, que encuentra en ella motivos para decir adiós al funcionalismo. La concepción funcionalista de las actitudes proposicionales exigiría una interpretación restringida de su contenido, pero la equivalencia entre actitudes proposicionales así entendidas y estados funcionales del organismo sólo puede mantenerse a costa de asentir a tres tesis inconsistentes: que las actitudes proposicionales tomadas por su contenido restringido son estados psicológicos, que los estados psicológicos son estados funcionales del organismo, y que dos instancias de un mismo tipo de estado funcional pueden diferir en cuanto al contenido restringido de la actitud proposicional que encarnan. Pero la última de estas tres tesis, que Baker dibuja como una de las ciudadelas cuya pérdida haría sucumbir al funcionalista, es en realidad tierra quemada que éste cedería sin titubeos: antes al contrario, lo que se afanaría en escudar es, como veremos, la idea de que no cabe diferencia semántica sin diferencia funcional cuando la semántica de los estados mentales, en su faceta restringida al menos, se pliega escrupulosamente a su imbricación en los patrones causales en los que se entreteja con estímulos, respuestas y otros estados mentales, exactamente como ocurre cuando comparamos la sintaxis de los procesos computacionales de un autómata con su interpretación semántica.

Vecina al cauce de esas preguntas acerca de si debemos siquiera dar cabida a los hechos del mundo en el seno de la explicación de la conducta mientras no cristalicen psicológicamente, o de cómo debamos describir los estímulos y las respuestas entre cuyos territorios se teje nuestra vida mental, fluye otra en torno al aparato conceptual que hemos de necesitar para dar cuenta de lo que queda en ese extraño istmo. Toda vez que la teoría de la identidad psicofísica había enfilado sus esfuerzos a la elucidación de la naturaleza de las sensaciones –respecto de las actitudes proposicionales se daba por bueno el análisis de Ryle–, el funcionalismo daría sus primeros pasos, también, mirando sólo de reojo al ámbito de las creencias y los deseos. Cuando éstos reclamaron la mirada, sin embargo, no tardó en hacerse patente que el hecho de que podamos albergarlos en una diversidad ilimitada, y de que cobijar unos nos obligue a dar abrigo también a otros, o al menos a poderlo hacer –esto es, la productividad y la sistematicidad que las actitudes proposicionales comparten con el lenguaje–, tal vez hiciese ilusorio entenderlos si no los pensábamos, precisamente, con la horma del lenguaje, articulando para ellos una semántica en la que, en definitiva, mantener una determinada actitud hacia tal o cual proposición pudiera ser tanto como encontrarse en un estado cuyas relaciones causales con otros estados internos que también constituyeran actitudes proposicionales reflejasen las relaciones semánticas entre las proposiciones pertinentes. Se iba bosquejando, entonces, la hipótesis del lenguaje del pensamiento a la que Fodor daría forma en 1975: un código simbólico en el que habían de llevarse a

cabo las operaciones computacionales postuladas por la teoría psicológica, una ***Lingua mentis*** que, a la luz de los severos aranceles que venían imponiéndose sobre el papel de las cosas mundanas en el ajetreo de la vida mental, no podía ser sino un **recinto umbrío**. Espoleada por la dificultad de concebir siquiera otra forma de engarzar creencias y deseos en la explicación psicológica sin recaer en el conductismo lógico, y por la confluencia con el trabajo en inteligencia artificial, la idea del lenguaje interior cobraría fuerza arrolladoramente. Que la mente lleva a cabo sus procesos computacionales en un lenguaje interno, que la explicación psicológica cotidiana en términos de creencias y deseos describe de hecho operaciones en dicho lenguaje, y que el contenido de creencias y deseos corresponde a las mismas expresiones del lenguaje interno con independencia de cuál sea el lenguaje natural del sujeto serían, mucho tiempo después, los dogmas de la teoría computacional de la mente que Putnam acabaría por impugnar.

Un estudio más minucioso de las razones –malas razones, a su juicio– que nos llevan a fiar nuestra comprensión de la vida mental a la existencia de ese lenguaje del pensamiento ha sido elaborado por Dennett: junto a las semejanzas ya apuntadas entre la estructura de nuestras creencias y deseos y la del lenguaje, a las que habría que añadir la posibilidad de asimilar la opacidad de lo mental a la intensionalidad de ciertas expresiones lingüísticas, o la subjetividad de lo mental a la presencia de elementos deícticos, convendría sopesar también las notorias dificultades, ya advertidas por Frege, que el desplazamiento de la noción de proposición hacia la de oración de un lenguaje interior nos permitiría eludir. Pero en la historia de la fascinación hacia esa *lingua mentis*, que habría sin duda de remontarse mucho más atrás, acaso hasta el mismo descubrimiento de esa voz con que uno habla para sí, sería preciso abordar también el relato de cómo el lenguaje interior fue desgajándose no ya de la lengua materna, sino del mundo que habitamos: de cómo, en otras palabras, pudo ir clausurándose aquella *camera obscura* que intuía Locke en el entendimiento para luego, resplandeciente bajo la luz que lograra colarse por los exiguos resquicios de los sentidos desde un mundo por lo demás enteramente oscuro para nosotros, dejarnos allí presos.

Pensar acerca de nuestras propias creencias y deseos bajo el prisma del lenguaje nos obliga más pronto o más tarde, como venimos viendo, a arrostrar la pregunta por el sentido en que la semántica interviene en el transcurso de la vida mental o en su precipitado en la conducta. El modo en que las propiedades semánticas de la *lingua mentis* se articularían con su sintaxis, de suerte que toda diferencia semántica estuviera codificada en una diferencia sintáctica, dejaba a la mano la conclusión de que la semántica sólo habría de figurar en la explicación psicológica de manera mediata, subrogada en la sintaxis. Ahora bien: esa conclusión, que encontraría una afortunada expresión en **La metáfora de la llave** –es la sintaxis de un deseo, no sus lazos con lo deseado, lo que lo dota de eficacia causal, como es la forma de una llave lo que provoca que abra una cerradura, no sus lazos semánticos con ella–, deja de nuevo en el aire la pregunta por **la soberanía del significado**, que es como decir por la irreductibilidad de la explicación psicológica.

La reflexión sobre los primeros autómatas alumbró ya la idea de que algunos de sus estados físicos pudieran operar como representaciones de aspectos del entorno, pero la relativa simplicidad de aquellos toscas maquinarias no abonaba la idea de que pudieran así constituir algo parecido a un lenguaje, susceptible de ser desgranado en una faceta sintáctica y una semántica –o, cuando lo hacía, como en los autómatas matemáticos, la presunta semántica venía nítidamente calcada de la de las expresiones que el ingenio estaba diseñado para calcular. Muchos menos, claro, tenía sentido la pregunta de si la eficacia causal de esas representaciones pudiera quedar ligada a nada distinto de las propiedades físicas de los estados en que quedaban materializadas. Pero cuando las computadoras digitales fueron desterrando a aquellos autómatas mecánicos no tardó en imponerse el enfoque lingüístico, y con él y la distinción entre propiedades sintácticas y semánticas –al fin y al cabo, aprestar a una de esas máquinas portentosas para tal o cual tarea requería dominar un lenguaje de programación. Ya en 1957, Allen Newell, Herbert Simon y J. Cliff Shaw repararon en que habían de tratar los símbolos y enunciados que intentaban codificar en su *Teórico Lógico* sin tomar en consideración su significado; Marvin Minsky, en 1958, describía la programación como un conjunto de procesos sintácticos sobre expresiones de naturaleza simbólica. El cognitivismo, así pues, heredaría una idea de lo sintáctico y lo semántico en la que confluían la noción lógica de gramática formal que, madurada en el ámbito de la teoría de la prueba, había fertilizado los primeros ensayos de simulación computacional del razonamiento y una noción de causalidad condensada en aquellas máquinas capaces imitar ciertas conductas de los organismos que aún poblaban los estantes de los laboratorios. Así, la idea de que la semántica de la vida mental se ceñirá sin descarríos a su sintaxis comenzaba ya adquirir el carácter fundacional que consagraría Fodor al formular la condición de formalidad que a su juicio pesaría sobre la relación entre los procesos psicológicos y las representaciones con las que operan. Explicar, entonces, las relaciones semánticas que aparecen entre nuestras creencias, deseos y conductas exigiría asumir la existencia de representaciones internas dotadas de unas propiedades sintácticas capaces de dotar de eficacia causal a la semántica de la vida mental: el lenguaje del pensamiento. De lo que se trataba, en suma, era ni más ni menos que de engarzar el orden de la racionalidad y el de la causalidad, y la sintaxis de la *lingua mentis* era la prodigiosa herramienta con la que habríamos de lograrlo. Entre esta concepción formalista de los procesos psicológicos y el solipsismo metodológico, por otra parte, hay lazos más estrechos de lo que pudiera parecer a simple vista: lo semántico es, precisamente, lo que concierne al vínculo entre la mente y las cosas, lo que quiebra la obediencia de las creencias o los deseos a la fisiología del organismo que los hospeda al uncirlos también a un mundo mucho más ancho.

Rastrear las raíces que la condición de formalidad hunde en el ámbito de la teoría de la prueba, ensanchando así la visión que del sustrato hilbertiano del funcionalismo se ha venido bosquejando, nos llevará a revisar también someramente el paulatino abandono de cierto logicismo ingenuo en la práctica de la investigación psicológica. Concurrían en alentar dicho tránsito, por un lado, la distancia entre el

comportamiento que cabría esperar a la luz de determinados cánones de racionalidad y el que de hecho mostraban los sujetos no ya en el tráfago de la vida cotidiana, sino incluso en los laboratorios de psicología experimental –como empezarían a hacernos ver Peter C. Wason o David Kahneman y Amos Tversky– , y, por otro lado, la inesperada dificultad para replicar artificialmente ciertas conductas que al desplegarse en nosotros con tanta naturalidad velaban lo arduo de su formalización –como tempranamente había denunciado Hubert Dreyfus. Este desencanto con el logicismo contribuiría poco a poco a acentuar las llamadas a la reivindicación del valor de la semántica que conviven en el seno del funcionalismo con el decidido énfasis formalista, llamadas que suelen avivarse cuando se hace preciso responder a la presión de algún argumento en favor del reduccionismo fisicalista aduciendo que atribuir contenido semántico a sus estados internos constituiría un trámite inexcusable si hemos de entender la conducta de ciertos organismos. Por regla general, esas exigencias de restitución de la semántica suelen terminar doblegándose al otorgarle en la explicación de la conducta, más o menos veladamente, un papel de orden pragmático, pero incluso en los trabajos de Fodor, como veremos, el gesto de recordar la preponderancia de la sintaxis y el de reclamar el peso de la semántica se suceden con desigual cadencia –convocarlos juntos es hacer palpable la contradicción.

Los vínculos entre el funcionalismo y la caracterización de la semántica de los estados mentales bajo parámetros internistas, que a su vez robustecen el primado de la condición de formalidad, han sido puestos en entredicho por Robert van Gulick. Es ilegítimo –argumenta van Gulick– encadenar al funcionalismo a una semántica internista como conclusión de razonamientos *a priori* acerca de la inercia causal de la semántica y el carácter local de la causalidad; sus objeciones, que arraigan en el pensamiento de Burge y Davidson, apelan al trecho lógico que media entre la naturaleza causal, *ergo* local, de una explicación y la de las propiedades en virtud de las cuales delimitamos los conceptos teóricos que intervienen en ella. Puesto que los puntales que sostienen el internismo como verdad de razón son débiles, la controversia habrá de dirimirse –concluye van Gulick– en la arena de los hechos: atendiendo a los hábitos de construcción de teorías de los propios psicólogos cognitivos. Pero en este ámbito, van Gulick coincidiría con Burge en advertir la pujanza de la apelación en todo su alcance a las propiedades semánticas de los estados mentales, tal como resulta del impulso que, a su entender, late de manera espontánea en la concepción representacional de la mente connatural al funcionalismo.

Así vista, la cuestión de la semántica se condensa en una pregunta por los compromisos ontológicos que conlleva cierto vocabulario teórico, y lo que queda por dilucidar es si la condición de formalidad estipulada por Fodor otorga a las propiedades semánticas de los estados mentales un papel en la explicación psicológica que resulte irreconciliable con una interpretación epifenomenista, o incluso eliminacionista, de su estatus ontológico –esto es, un papel que desborde el orden de su valor heurístico o pragmático–, o, en caso contrario, si cabe alguna

reformulación de la relación entre la sintaxis y la semántica de los estados mentales que conjure cualquier devaluación ontológica y que no pase por asumir, con Burge, una plena renuncia al individualismo. Ese sendero, que tantearemos de la mano de Heil, nos llevará de vuelta a la tesis de que evitar los taludes que nos abocarían al eliminacionismo –por tanto, piensa Heil, a la incoherencia–, y de cuyo filo, aunque no nos demos cuenta, caminamos tan cerca si nos aventuramos en las vastas regiones del externismo como si preferimos quedarnos cerca de casa, en los apriscos del internismo, exige repensar nuestra concepción de la causalidad: su propuesta, en particular, es que debemos dotarnos de los aparejos precisos para integrar en ella la idea de sobredeterminación. Pero que los efectos de albergar tal o cual creencia o tal o cual deseo puedan estar plenamente determinados por su semántica a la vez que por su sintaxis, aunque acaso elimine algunas de las motivaciones para abrazar el eliminacionismo o el epifenomenismo, no nos alcanza para afianzar la irreductibilidad de la explicación psicológica que buscábamos en la soberanía del significado. Si hemos de cosechar ese fruto, habrá de ser porque hayamos encontrado, como han sabido ver Dretske, Fodor o Braddon-Mitchell y Jackson, una noción de la causalidad que nos permita conceder eficacia causal al contenido semántico de los estados mentales *qua* contenido semántico, desasida de la que la estructura sintáctica de las oraciones de la *lingua mentis* en la que se encarnan pueda dispensarles. No venimos a decir, en suma, mucho más de lo que Sócrates dejara dicho sobre el modo en que Anaxágoras de Clazómenes recurría al *Nous* a la hora de explicar las acciones de los mortales: que un hombre beba de su propia mano la cicuta en virtud de las creencias sobre la justicia que se ha forjado es irrelevante si lo mismo habría hecho en virtud de la disposición de sus huesos, músculos y tendones.

El lugar donde Braddon-Mitchell y Jackson creen haber hallado la clave que eludiera a Anaxágoras vendría a ser éste: podremos reconocer en las creencias de Sócrates sobre la justicia la causa de que éste repudie los ruegos de Critón de que acepte huir de Atenas si dichas creencias no sólo causan su comportamiento en virtud de la actividad nerviosa en que cristalizan, sino que tal cosa forma parte del patrón de relaciones funcionales que nos permite clasificarlas precisamente como creencias sobre la justicia. Con otras palabras: si por medio de su encarnación fisiológica una creencia causara una conducta ajena a su definición funcional como tal creencia, la creencia sería la causa de la conducta, pero no lo sería *qua* creencia, es decir, *qua* creencia con tal contenido semántico. Regresamos, así, a una problemática –la de las **Cadenas causales díscolas**– que ya en 1966 dejó descrita Chisholm en su vigoroso intento de desbaratar cualquier reconstrucción del concepto de acción humana que pretendiera esquivar lo teleológico. De manera parecida a como ocurría en aquellos análisis que Chisholm censurase, la dificultad para Braddon-Mitchell y Jackson reside, claro, en justificar la exclusión de tal o cual conducta de la madeja de causas y efectos que definen a tal o cual estado mental sin apelar al papel del contenido semántico de lo mental del que pretenden estar rindiendo cuentas: no es sencillo, de nuevo, preservar la relevancia explicativa de lo mental con el mismo gesto con que se deniega su eficacia causal. El intento nos llevará pronto de estas

indóciles cadenas causales en las que lo mental se hace presente mostrando su propia debilidad a la necesidad de admitir en el cuerpo teórico de las ciencias especiales generalizaciones **y leyes *cæteris paribus***, que especifiquen sólo de forma incompleta las condiciones en que se impondrán.

Entre tantos otros, el reto de instaurar un espacio de causalidad que no sea hostil a la semántica ha sido afrontado también por Robert Cummins, quien confía en encontrarlo recurriendo a las mismas herramientas que permitieron forjar el delicado equilibrio en que la concepción funcionalista de la mente pervive entre fisicalismo y conductismo. De lo que se trata, entonces, es de hallar generalizaciones sobre los efectos de determinados tipos de estados mentales que se mantengan en pie en virtud de las propiedades semánticas de las instancias que conforman dichos tipos, no en virtud de sus propiedades sintácticas –ni, como en Braddon-Mitchell y Jackson, en virtud de sus propiedades de ambos órdenes–, de suerte que expresar las regularidades en cuestión mediante un vocabulario parcamente sintáctico nos forzase a un alambicado y arbitrario edificio disyuntivo. Realización múltiple, en suma, de la semántica en la sintaxis, en el bien entendido de que esto implica tanto la existencia de generalizaciones que involucren la adscripción de contenido semántico estados mentales como la vulneración de la condición de formalidad que sometería esas generalizaciones al yugo de otras, parejas, de índole sintáctica. Pero dar por supuesto que entre el significado de los estados mentales y la forma de sus expresiones en un hipotético lenguaje del pensamiento pueda calcarse la relación entre los estados mentales y su fisiología conduce, como veremos, a que la subordinación de semántica a la sintaxis acabe convirtiéndose en un baldío enfrentamiento de intuiciones contrapuestas, o acaso de intuiciones y estipulaciones.

Una vez más, entonces, la controversia se agosta en el terreno de la causalidad. Revestir de eficacia causal a lo mental, o a su faceta semántica, no es otra cosa que encontrar leyes causales en las que figure, pero queda abierta la pregunta de si habría de tratarse de leyes causales en las que figure y que no podamos reducir a otras en las que no lo haga, o si ese privilegio está reservado a propiedades físicas elementales capaces de trabar relaciones mecánicas con otras de su misma naturaleza. Mientras no hallemos el modo de destensar el entramado que liga la aparición en leyes causales de cualquier propiedad de orden superior a esa eficacia primitiva, cada intento de redimir la noción de causas mentales desemboca inexorablemente, si no en la contradicción, en declaraciones de soberanía explicativa de la psicología que se diluyen, como venimos viendo, en la rendida súplica de una incierta relevancia a efectos pragmáticos. Como han señalado Frank Jackson y Philip Pettit, las mismas fuerzas que empujarían a un dualismo interaccionista de inspiración cartesiana hacia las ciénagas del epifenomenismo operan también, pese a la observancia materialista de que el funcionalismo hace gala, sobre su empeño por preservar un coto inexpugnablemente psicológico en nuestra comprensión del mundo. En el convencimiento de que no se salva dicho coto ni por adjudicar la eficacia causal de los estados mentales, o de su contenido semántico, a unas propiedades fisiológicas enteramente opacas para la psicología coloquial que nos incita a tomarlos como

causas, ni por otorgársela a ellos mismos por la mera razón de que guarden determinados vínculos con esas propiedades fisiológicas, Jackson y Pettit intentarán hacer germinar la relevancia explicativa del vocabulario psicológico asemejando las propiedades funcionales a disposiciones, y armando una esmerada defensa, en la que la idea de realización múltiple vuelve a resultar crucial, de la tesis de que la relevancia causal de una disposición no se agota en la de su fundamento categórico. Pero, tan reacios a un cuestionamiento de la noción mecánica de causalidad como sabedores de que si no viene acompañado de ese cuestionamiento todo el esfuerzo por rehuir el epifenomenismo acaba más pronto o más tarde desnudando su futilidad, Jackson y Pettit preferirán argumentar que cuantas razones podamos recoger para dudar de la eficacia causal de las propiedades psicológicas las tenemos también respecto de cualesquiera otras, salvo las propiedades físicas elementales, que encontremos ya en la ciencia ya en cualquier otra vertiente de nuestra comprensión del mundo: ésa, se diría, es para Jackson y Pettit la carga de desencanto que no tendremos más remedio que sobrellevar.

A muy parecida conclusión ha llegado Fodor, que, sin embargo, sospecha que sean cuales sean los argumentos que nos hayan convencido de que no hay genuinas causas ni efectos en el mundo que cotidianamente habitamos, algo debe de haberse torcido en su transcurso. Desandar los pasos que nos han traído a resignarnos con esa visión de una realidad enteramente inerte ante nuestros ojos, cuyo interminable trasiego de causas y efectos se despliega siempre a escondidas, pasaría –argumenta Fodor– por ensanchar el espacio que la idea de realización múltiple nos concede pero no, como pretendía Cummins, aplicándolo a la relación entre semántica y sintaxis, sino mediante un análisis del papel de las cláusulas *cæteris paribus* en la explicación científica. Como veremos, la táctica de Fodor viene en realidad precipitada por las críticas de Schiffer a la propia noción de ley *cæteris paribus*: si a juicio de Schiffer la indefinición que introduce la cláusula *cæteris paribus* anula el carácter legaliforme del enunciado al que acompaña, para Fodor, por el contrario, lo que hace la cláusula *cæteris paribus* es aludir de forma general a las condiciones adicionales cuya conjunción con cierta encarnación física de un determinado estado funcional resulte suficiente para que se dé cierta conducta, sin que ninguno de los términos de la conjunción se baste para producir el mismo resultado. Así, podríamos distinguir tres tipos de leyes, o, si se quiere, de enunciados con apariencia de ley: las leyes estrictas que proveen los fundamentos de la física, las leyes *cæteris paribus* de las ciencias especiales, y las generalizaciones vacías que, como denuncia Schiffer, ni son genuinas leyes ni deben formar parte del discurso científico. La clave –claro está– se halla en poder diferenciar nítidamente las leyes *cæteris paribus* de esas pseudogeneralizaciones: la propuesta de Fodor es que las leyes *cæteris paribus* pueden admitir excepciones –cuando no se cumplen las condiciones adicionales– e incluso excepciones absolutas –cuando no es posible fijar dichas condiciones adicionales con más precisión que la que proporciona la propia cláusula *cæteris paribus*–, pero no admiten, a diferencia de las generalizaciones vacías, excepciones generalizadas –es decir, excepciones absolutas que afectan tanto a la propia ley como

a cualquier otra en la que aparezca como antecedente el mismo estado funcional. Pero dado que los estados internos que figuran en ellas vienen definidos en términos funcionales, las leyes psicológicas no pueden constituir generalizaciones vacías. Antes bien, el hecho de que un determinado estado físico no acarrease las consecuencias que fija la definición funcional de un estado psicológico en ninguna de las leyes en que dicho estado psicológico pueda aparecer como antecedente simplemente nos mostraría que ese estado físico no es una encarnación del estado psicológico en cuestión. Así, no sólo la relevancia explicativa del vocabulario psicológico habría quedado afianzada, sino también su irremplazabilidad: lo que las cláusulas *cæteris paribus* nos proporcionarían –por lo demás, como en otras ciencias especiales– sería la posibilidad de cuantificar sobre los mecanismos subyacentes a una regularidad causal cuando su naturaleza física resultara heterogénea. En definitiva, aun cuando *explicar* en tanto que *dar cuenta de una relación causal singular* pudiera hacerse sin recurso a otro vocabulario que el de la física, *explicar* en tanto que *apresar una generalización sobre relaciones causales* exigiría las herramientas de las ciencias especiales. O, dicho de otro modo, aunque los mecanismos que implementan en cada caso particular las leyes de la psicología sean de naturaleza sintáctica –en último término, física–, dichas leyes son semánticas hasta la médula. Pero a la hora de examinar si el análisis de Fodor de las cláusulas *cæteris paribus* puede apuntalar una defensa robusta de la autonomía explicativa de la psicología, o si por el contrario se ve aquejada de la misma inestabilidad que hemos visto provocar que intento tras intento de restaurarla se venciera del lado de una autonomía sin irreductibilidad, veremos agigantarse una vez más la sombra de la renuncia a uno de los pilares de nuestra concepción de la causalidad: la idea de poderes causales según la cual los que quepa atribuir a una propiedad de orden superior se agotan en los de las propiedades físicas en que se materializa.

Sobre la valoración del peso de la semántica de los estados mentales en el modelo de explicación psicológica auspiciado por el cognitivismo confluyen **Nociones de lo sintáctico**, y acaso también de lo semántico, provenientes de muy distintos ámbitos: de un lado, como quedó anticipado, el empeño de formalización del **pensamiento** lógico que cobra ímpetu en la teoría de la prueba, **y** de otro, la larga tradición de estudios estructurales del **lenguaje** que germina en la lingüística moderna.

Convendrá entonces tomar un poco de perspectiva escudriñando cómo cierta idea de la relación entre la sintaxis y la semántica va madurando en el marco de los primeros esfuerzos de construcción de autómatas inteligentes, netamente mecánicos primero y regulados por mecanismos de índole computacional, en los se destila todo la labor de formalización del cálculo lógico, después. La controversia reciente sobre el papel de la semántica y la condición de formalidad se reconoce en forma seminal, como se ha anticipado, en la encendida polémica que en las primeras décadas del S.XX, a cuenta de la legitimidad del vocabulario teleológico en la explicación de la conducta de los organismos o de autómatas que la remeden, enfrentase a Loeb y a Jennings. En el pensamiento de Ralph B. Perry encontraremos ya en 1917 una

meditación matizada y madura sobre la diferencia, en cuanto a compromisos ontológicos, entre admitir dicho vocabulario a título de punto de vista sobre ciertos fenómenos y declararlo epistemológicamente irremplazable; a través del propio Perry y de Nicolas Rashevsky la cuestión pronto impregnaría los círculos de Tolman y Hull. A la par que se iban fraguando estas reflexiones acerca de lo teleológico, cobraba también relieve la naturaleza semántica de ciertos conceptos empleados en la explicación de la conducta de los autómatas: si en 1925, Alfred J. Lotka aludía a los estados internos del escarabajo mecánico que había construido como una representación del entorno –crudamente correlacional, eso sí–, en 1943, en Craik, encontraremos ya esos vínculos correlacionales acompañados de un aparato inferencial que capacita al autómata para extraer de ellos conclusiones acerca de su entorno mediante operaciones simbólicas, procurándose así información que desborda la que le facilitan sus dispositivos, digamos, sensoriales. Poco después, la intuición ya bosquejada en Jennings de que a la hora de explicar la conducta de autómatas u organismos complejos no tendríamos más remedio que asumir un nivel de abstracción que nos apartase del vocabulario de la física o la química comenzaría a hacerse fuerte merced al desarrollo de los lenguajes de programación de alto nivel, que alejaban la tediosísima y poco fecunda tarea de codificar cada proceso en términos de posiciones de conmutadores eléctricos –esto es, en código máquina. En el trabajo fundacional de aquellos pioneros de la programación –tanto en el de Newell, Shaw y Simon como en el de McCarthy– reconoceremos sin dificultad la huella de Hilbert y su apuesta por fiarlo todo a las características formales de la representación del problema, de modo que la semántica venga fielmente detrás. Ya entonces, sin embargo, la inevitable prevalencia *ante litteram* de la condición de formalidad en la práctica de la programación convivía con un intento decidido de diferenciar aquella incipiente disciplina de la investigación en inteligencia artificial respecto de un deslustrado conductismo apelando al uso de vocabulario mentalista –teleológico, semántico–; así lo constataremos, por ejemplo, en el pensamiento de Minsky.

Más hondas aún en términos históricos –Port-Royal, 1660– se hallan las semillas de una concepción de la sintaxis en la que el análisis de las estructuras oracionales comunes a cualquier lenguaje natural se vale de un conjunto de categorías, articuladas en torno a las de *sujeto* y *predicado*, cuya pretendida universalidad descansa sobre sus lazos con las categorías lógicas que articularían el pensamiento. Si bien esas categorías sintácticas se encuentran estrechamente entrelazadas con nociones semánticas, no cabe duda de que el intento de destensar esos vínculos ha constituido una de las corrientes que más vigorosamente han agitado las aguas de la lingüística desde entonces. Veremos, no obstante, que la defensa de la autonomía de la sintaxis emprendida por Chomsky no implica, ni mucho menos, la verdad de la tesis de que toda diferencia semántica tenga codificación sintáctica, y que si bien la eliminación de lo semántico no es para la lingüística, como tampoco para la psicología, un escenario impensable hoy, la reivindicación del papel irreductible que habría de corresponderle en una

comprensión madura del lenguaje, así como de su inevitable entrelazamiento con la sintaxis, son posiciones no menos vigentes en el paisaje metateórico de la lingüística.

Desde luego, encontrar diferencias semánticas entre oraciones que no se acompañen de diferencias sintácticas es sencillo en el ámbito del lenguaje natural, tan fácil como encontrar ambigüedades sintácticas que contaminan la interpretación semántica de un enunciado. Pero el ejercicio es trivial, pues, de pretender desarmar empíricamente la condición de formalidad, donde tendríamos que encontrar diferencias semánticas sin reflejo sintáctico, o ambigüedades sintácticas que comporten secuelas semánticas, sería en el singular reducto de la *lingua mentis*. Será preciso, así pues, fijar unos criterios que nos permitan dirimir si, pongamos por caso, una determinada ambigüedad sintáctica detectada en la expresión de una creencia en lenguaje natural tiene su origen en la estructura de dicha expresión o bien en la propia representación mental hacia la que el sujeto, de acuerdo con las tesis funcionalistas, mantiene la actitud de creer. Quizá el único contexto en que pudiéramos concluir que la ambigüedad es de raíz cognitiva y no lingüística sea aquel en que el sujeto asintiera explícita o implícitamente a cada una de las diversas interpretaciones de la oración que expresa una creencia, sin que ocurriera otro tanto cuando se viera expuesto a una versión desambiguada de la misma.

Pero si hablamos de ambigüedades semánticas, en cambio, la posibilidad de diferenciar entre las propiedades de la oración que expresa un estado mental y las del propio estado mental se difumina. Las marcas de intensionalidad presentes en el lenguaje con empleamos para expresar nuestras creencias y deseos se reflejan límpidamente en el papel de esas creencias y deseos en la determinación de la conducta: el hecho es que uno actúa a tenor de lo que cree o desea con independencia de que esas creencias o deseos se refieran o no a lo que uno da por hecho que se refieren, o se refieran a algo en absoluto, etc. Eso, que sustenta las intuiciones solipsistas que hemos reconocido en el núcleo del cognitivismo, contribuye también a tensar los lazos entre la semántica de la expresión lingüística de una actitud proposicional y la de su expresión en la *lingua mentis* de tal modo que subordinar ésta a una sintaxis estrictamente formal desprovee a aquélla de todo margen de soberanía. El impulso fundacional que la psicología cognitiva toma de la noción de representación interna se desdobla así en el convencimiento de que la conducta no depende tanto del entorno del organismo como del modo en que éste se represente dicho entorno; de esta manera, lo que podían parecer a primera vista tendencias contradictorias en el seno del cognitivismo –una cierta reivindicación de la semántica, sedimentada en el decisivo cometido que se confiere en la explicación psicológica al modo en que nos representamos las cosas, frente a la insistencia en que la semántica, entendida como las cosas que de ese modo nos representamos, es inerte en la determinación de la conducta– resultan ser después de todo facetas de una única idea. Los compromisos internistas que marcan los primeros tiempos del funcionalismo aparecen desde esta perspectiva como consecuencia natural del papel que la confrontación con el conductismo desempeña en su génesis.

Ahora bien, la imposibilidad de dar con diferencias semánticas que además de carecer de reflejo sintáctico en la representación de un estado mental en el lenguaje del pensamiento conlleven repercusiones de orden conductual obedece en parte a motivos espurios que será preciso aislar. Veremos, en efecto, como en la concepción funcionalista de la sintaxis y la semántica trasluce por doquier una identificación como sintáctico de todo aquello que comporte efectos en la conducta, o en el mismo despliegue de la vida mental, que cobra valor de *fiat*. No son propiedades sintácticas todas las propiedades físicas de un sistema –la sintaxis, se entiende, es alguna suerte de abstracción sobre propiedades físicas–, y los criterios en virtud de los cuales se distingue lo sintáctico incluyen con frecuencia precisamente la eficacia causal de cara a la regulación del comportamiento de un sistema. Pero si hacemos de una determinada propiedad de una representación mental una propiedad sintáctica en la medida en que cabe atribuirle consecuencias en la conducta, es tan obvio como poco interesante que luego no lograremos encontrar propiedades de las representaciones mentales que acarreen tales consecuencias sin ser sintácticas –por ejemplo, propiedades semánticas. Acaso, entonces, como ha argumentado Cummins, la impresión de que en las propiedades semánticas de una representación mental resultan inertes en cuanto a la determinación de la conducta repose a fin de cuentas en una deficiente comprensión de los vínculos entre sintaxis y semántica, contra la que podrías servir de antídoto recordarnos que sin semántica no hay representación ni, por tanto, sintaxis, y que *formalizar* la relación entre ambas no es otra cosa que forzar que toda diferencia semántica tenga un reflejo en la sintaxis. Acaso también el artificioso destierro de la semántica que se deriva de todo esto haya acabado instigando su contrabando bajo el embozo de la sintaxis, de igual manera –veremos– que diversos conceptos mentalistas terminaban por infiltrarse en los análisis disposicionales y los modelos de aprendizaje de los conductistas. Por otra parte, desligar por completo la semántica de nuestras creencias y deseos de la sintaxis de los cómputos internos y tomar a esta última, siguiendo el consejo de John McDowell, por una mera condición de posibilidad de aquélla, como pudiera serlo cualquier constante física, nos aboca –lo veremos despacio– a un dilema entre la indigencia de la semántica y su dispensabilidad que nos deja lejos del engranaje entre causas y razones que buscábamos.

Con la controversia en torno a la eficacia causal de la semántica nítidamente enclavada en el seno de las tensiones entre una lectura realista de los conceptos básicos de la psicología coloquial y la tentación de darlos por abolidos, nos detendremos finalmente en sopesar **Un ensayo de restitución** de su valor epistemológico y su robustez ontológica que aspira a esquivar la ruptura con la interpretación internista de la semántica a la que ese realismo psicológico parece en alguna medida abocado. La tesis fundamental sobre la que Michael Devitt ha articulado dicho ensayo es la de que la noción de contenido restringido aporta una riqueza semántica mucho mayor de la que es aparente *prima facie*, y, en particular, hace factible una reconstrucción exhaustiva y fiel de los rasgos veritativo-funcionales de las propiedades plenamente semánticas que adjudicamos al contenido amplio:

dicho de otra manera, el contenido restringido es contenido proto-veritativo-funcional. De lo que se trata entonces es de mostrar que dicha noción de contenido restringido es tanto necesaria –no basta con la sintaxis– como suficiente –no es preciso recurrir al contenido amplio propugnado por Burge– de cara a la teorización psicológica. En su defensa de la necesidad del contenido restringido Devitt descansa, como se analizará detenidamente, en una escrupulosa circunscripción de la idea de propiedad sintáctica que la depura de la petición de principios respecto de la eficacia causal sobre la que, según venimos de ver, parece radicar el primado de la sintaxis en ciertas concepciones del cognitivismo: bajo esa demarcación, la posibilidad de invocar los lazos con estímulos y conductas en sentido proximal que nos proporciona la noción de contenido restringido, pero que se está vetada a las propiedades sintácticas de una representación mental, se vuelve imprescindible si hemos de comenzar siquiera a armar una teoría psicológica. La defensa que Devitt, por otro lado, plantea de la tesis de que no es necesario que nos adentremos más allá del contenido restringido, hacia una semántica plenamente referencial, no tiene más fuerza que la apelación a las consabidas intuiciones solipsistas acerca del parapeto que nuestra representación del mundo impone entre las cosas y nuestra reacción a las cosas –intuiciones que, bien lo sabemos ya, cabe cuestionar. En la práctica, además, veremos como el propio Devitt tiende a introducir estados mentales identificados por su contenido amplio en los ejemplos de posibles leyes psicológicas que salpican sus argumentos. Su ensayo de restitución de una semántica internista para la psicología, en suma, sólo puede considerarse inconcluso.

Las preguntas para las que venimos tratando de encontrar respuesta, o al menos de perfilar con nitidez lo que nos aparta de haberlas encontrado, podrían ser el fruto malogrado de preocupaciones filosóficas comprensibles pero espurias –así ha tratado tenazmente de hacérnoslo entender McDowell–: deberíamos, entonces, desistir de la tarea y aprestarnos a entender cómo llegamos a creer que ésta era practicable, o incluso apremiante. Sólo después, cuando pudiéramos plantear las preguntas en otros términos –en particular, si hemos de seguir los pasos de McDowell, destejiendo los lazos que damos por hecho existen entre los sentidos y la pasividad, el juicio y la espontaneidad–, podríamos también responderlas en otros términos: sencillamente encogiéndonos de hombros o, a lo sumo, haciendo ver que esos rasgos singulares de nuestra vida mental que provocaron aquellas primeras, torcidas preguntas –la espontaneidad del juicio, la intencionalidad, la semántica, la teleología– son a un tiempo naturales y no naturales, en la medida en que pertenecen a nuestra *segunda naturaleza*. Pero –concluiremos– por rico que sea el entramado conceptual en que nos sitúa la idea de segunda naturaleza, ni podrá mitigar por sí solo la perplejidad que despierta constatar que vivimos a un tiempo en el mundo de las causas y en el de las razones –que transitamos silenciosamente entre uno y otro, que somos, diría Emilio Lledó, **Naturaleza en la naturaleza**– ni podrá acallar las preguntas, que volverán a avivarse en su seno.

MUNDO, PALABRA, MENTE

El genio de la lámpara

En las primeras páginas de *Los tónicos de la voluntad*, antes de adentrarse en el escrutinio de las cualidades que debe afanarse en ejercitar el investigador científico y las debilidades que pueden apartarle de su tarea, Santiago Ramón y Cajal aborda una reflexión metodológica de índole general, proveniente de su discurso de ingreso en la Real Academia de Ciencias Exactas, Físicas y Naturales, en 1897. Enarbolando así su ánimo de avivar "[…] el entusiasmo de la juventud estudiosa hacia las empresas del laboratorio" (Ramón y Cajal 1897/1941: 17), Cajal rechaza resignarse, por prematuros, tanto al fenomenalismo positivista propugnado por Claude Bernard –que glosa como la tesis de que "[…] el investigador no puede pasar del determinismo de los fenómenos, su misión queda reducida a mostrar el cómo, nunca el porqué, de las mutaciones observadas" (Ramón y Cajal 1897/1941: 28)–, como al rendido agnosticismo al que había dado cuerpo Émil du Bois-Reymond:

> Frente a los enigmas del mundo material, el investigador de la naturaleza está habituado desde hace tiempo, con viril renuncia, a pronunciar su *ignoramus...* donde él ahora no sabe, pero podría acaso saber, o sabrá un día, en ciertas condiciones. Pero frente a los enigmas relativos a qué sean materia y fuerza y cómo ellas puedan ser capaces de pensar debe, una vez por todas, plegarse a un veredicto mucho más duramente renunciatorio: ¡*ignorabimus*! (du Bois-Reymond 1872 *apud* Bueno 1990: 69)

Quién sabe si pesaban ya entonces en el ánimo del abnegado fisiólogo prusiano las palabras que había dejado escritas su buen amigo John Tyndall –un reputadísimo físico nacido en una pequeña aldea del sureste de Irlanda, aunque educado científicamente en Marburgo–, que se mostraba convencido de nuestra completa invalidez a la hora de comprender la relación entre la consciencia y la actividad cerebral. En el verano de 1868, Tyndall había pronunciado en Norwich, ante la Sección de Física y Matemáticas de la *British Association for the Advancement of Science*, una conferencia en torno al "Alcance y límites del materialismo científico" que después formaría parte de sus populares *Fragments of Science for Unscientific People* (Tyndall 1871). A ojos de Tyndall, el abismo entre lo mental y lo material se revelaba insalvable:

> […T]he passage from the physics of the brain to the corresponding facts of consciousness is unthinkable. Granted that a definite thought, and a definite molecular action in the brain occur simultaneously; we do not possess the intellectual organ, nor apparently any rudiment of the organ, which would enable us to pass, by a process of reasoning, from the one to the other. They appear together, but we do not know why. Were our minds and senses so expanded, strengthened, and illuminated, as to enable us to see and feel the very molecules of the brain; were we capable of following all their motions, all their

> groupings, all their electric discharges, if such there be; and were we intimately acquainted with the corresponding states of thought and feeling, we should be as far as ever from the solution of the problem, "How are these physical processes connected with the facts of consciousness?" The chasm between the two classes of phenomena would still remain intellectually impassable. Let the consciousness of love, for example, be associated with a right-handed spiral motion of the molecules of the brain, and the consciousness of hate with a left-handed spiral motion. We should then know, when we love, that the motion is in one direction, and, when we hate, that the motion is in the other; but the "Why?" would remain as unanswerable as before. (Tyndall 1871: 86-87)[1]

La *British Association* era ya entonces, pese a su juventud, una institución célebre, a cuenta sobre todo del encendido debate acerca de la evolución y el origen del hombre que en 1860 había enfrentado en su seno, en Oxford, a Thomas H. Huxley y al obispo Samuel Wilberforce. Precisamente Huxley, en la primera edición de sus *Lessons in Elementary Physiology*, había dado muy viva expresión a la misma perplejidad sobre la que se explayaría Tyndall –a ambos les unía también una amistad que se remontaba a la juventud:

> [H]ow is it that anything so remarkable as a state of consciousness comes about as a result of irritating nerve tissue, is just as unaccountable as the appearance of the Djin when Aladdin rubbed his lamp. (Huxley 1866: 193)

Sólo en lo que quedaba de siglo, las *Lecciones* de Huxley conocerían treinta nuevas ediciones o reimpresiones[2]. En la tercera edición, la metáfora del genio de la lámpara que abría la *Lección VIII* –"Sobre las sensaciones y los órganos sensoriales"– había dejado paso a la idea, más aséptica, de que la consciencia parecería contarse entre las propiedades últimas de la naturaleza[3]. La claridad con la que aparece entonces la

---

[1] La evocación de Leibniz (1715: §17) es vívida, y está también presente en los argumentos de du Bois-Reymond –*cf. infra*:

> Por otra parte, nos vemos obligados a confesar que la *percepción* y lo que depende de ella es *inexplicable por razones mecánicas*, es decir, mediante las figuras y los movimientos. Y si imaginamos una máquina cuya estructura haga pensar, sentir, tener percepción, se la puede concebir de mayor tamaño conservando las mismas proporciones, de manera que se pueda entrar en ella como en un molino. Concedido esto, al visitarla por dentro sólo se hallarán piezas que se empujan unas a otras y jamás algo con lo que explicar una percepción.

[2] Sólo doce consigna Tennant (2007) en su certero estudio de las raíces que la reflexión contemporánea sobre lo mental hunde en la tradición de la *Naturphilosophie* alemana de mediados y finales del s. XIX. La detallada bibliografía de Huxley que se incluye como apéndice al tercer volumen de *The Life and Letters of Thomas Henry Huxley*, al cuidado de su hijo Leonard (Huxley 1900: 214), recoge una por una las treinta que sucedieron a la edición original, de 1866.

[3] Como sagazmente ha anotado Tennant, es díficil averiguar los motivos precisos que llevaron a Huxley a desterrar al genio, si bien cabe pensar que pudo caer en la cuenta de que "[…] some readers had gained either the mistaken impression that he was inclined to take seriously the possibility of supernatural phenomena such as Djins […], or […] the mistaken impression that Huxley 'supported some divine intervention in the creation of consciousness'" (Tennant 2007: 750-751) –una idea, por cierto, que Alfred Wallace había popularizado en el seno de la controversia sobre el origen del hombre

tesis de que no cabe explicación del hecho de la consciencia hace que, aun a pesar de que lo hiciera tratando de aclarar sus propios sentimientos religiosos, no deje de resultar significativo el hecho de que fuese precisamente Huxley quien acuñó la noción de agnosticismo:

> We class sensations, along with *emotions*, and *volitions*, and *thoughts*, under the common head of *states of consciousness*. But what consciousness is, we know not; and how it is that anything so remarkable as a state of consciousness comes about as the result of irritating a nervous tissue, is just as unaccountable as any other ultimate fact of nature. (Huxley 1866/1872: 188)

Esa inexplicabilidad de la consciencia encuentra su eco quizá más inquietante en la imposibilidad de determinar su presencia, de averiguar si otros seres, distintos de nosotros mismos, están dotados siquiera de lo que McGinn (1999), enardeciendo el tono agnóstico de sus precursores, bautizaría como "la misteriosa llama"[4]. Tal era, a ojos de Huxley, un corolario ineludible del hecho de que no podamos dar cuenta del surgimiento de la consciencia. En los párrafos finales de una breve alocución que dictara ante la efímera *Metaphysical Society*, en otoño de 1870, Huxley, tras discutir la evidencia experimental que la sección a distintos niveles de las fibras nerviosas de un sapo pudiera aportar a la controversia sobre la naturaleza extensa o inextensa, divisible o indivisible del alma, apunta a modo de *caveat*:

> I have not attempted to discuss the question whether the soul of the frog possesses consciousness, because this appears to me to be a totally insoluble problem.

---

que finalmente, como es sabido, forzaría a Darwin (1871) a hacer explícitas sus tesis al respecto, una controversia en la que Huxley había cobrado especial protagonismo debido a su acre enfrentamiento con Wilberforce.

[4] O, *a fortiori*, de si las experiencias conscientes de dichos seres, cualesquiera que fueran, son cualitativamente semejantes a las nuestras o, acaso, de todo punto incomparables. Ni siquiera, argumentaría du Bois-Reymond (1898: 42-43), nos es dado saber a ciencia cierta si el efecto de un estímulo sobre un organismo es placentero o doloroso, *i.e.*, semejante a nuestra experiencia subjetiva de placer o a la de dolor. Preocupaciones recientes, como la que atañe a la imposibilidad de conocer, por así decir, fenomenológicamente modalidades sensoriales de las que carecemos, como la ecolocalización (Nagel 1974), o a la de descartar racionalmente la existencia de espectros sensoriales invertidos (Shoemaker 1982, Block 1990b, pero *cf*. Locke 1689: II, XXXII; §15), arraigan también, según ha mostrado Tennant (2007), en el sustrato de este agnosticismo científico en torno a la consciencia que caló en la fisiología decimonónica. Como atinadamente señala también Tennant (2007: 760), no sería extraño que el vínculo entre McGinn y du Bois-Reymond tuviera un eslabón en Herbert Feigl, un ardiente defensor de la identidad psicofísica que acabaría escorándose hacia el eliminacionismo, cuya influencia en el pensamiento anglosajón del s. XX, avivada por su trabajo como coeditor, junto con Wilfrid Sellars, de *Philosophical Studies*, fue enorme. No en vano, Feigl se había formado en el círculo de Rudolf Carnap en Viena, quien había reconocido su deuda con du Bois-Reymond en lo relativo a la formulación del problema de la mente y el cuerpo (Carnap 1928: 266-267). En su influyentísima monografía sobre *Lo 'mental' y lo 'físico'*, Feigl se muestra convencido de que las relaciones psicofísicas –la correlación Ψ-Φ, en su condensada notación– son irreducibles desde una perspectiva científica: "There is no plausible *scientific* theory anywhere in sight which would explain just why phenomenal states are associated with brain states" (Feigl 1958: 105).

> Every one will discover, if he considers his own actions, that he is constantly performing operations directed towards special ends of which he has no consciousness whatever. And therefore it must be granted that it is possible that all the far less complex actions of the frog *may* be equally devoid of consciousness. Whether they are so or not, is a point on which no positive evidence is attainable, or even conceivable. (Huxley 1870: 7)

La idea de que ni siquiera podemos concebir qué clase de evidencia pudiera aducirse para legitimar la atribución de consciencia a otros organismos queda así estrechamente ligada a la de que el hecho mismo de la consciencia sea de tal naturaleza que no podamos rendir cuentas de él. Poco antes, en su conferencia de 1868, Tyndall –recordemos– sea había mostrado igualmente tajante: la transición de la fisiología a la consciencia es "impensable" (Tyndall 1871: *supra*). Apenas unas líneas más abajo, el propio Tyndall deja labrada una metáfora que, desprovista de los ribetes mágicos de la lámpara de *Las mil y una noches*, gozaría de mayor longevidad: entre la fisiología y la consciencia se abre un abismo, una sima que no somos racionalmente capaces de sobrepasar. El abismo –*chasm*– de Tyndall sería pronto un golfo –*Klüft*– en la pluma de du Bois-Reymond, un espacio "sobre el que no hay puentes ni alas que puedan llevarnos, pues nos alzamos frente a los límites de nuestro ingenio" (du Bois-Reymond 1898: 33). Sobre cuán sospechoso resulta lo inerme que ese vacío se muestra en nuestra cotidianeidad –indicio, a su entender, de que no es más que un artificio ocasionado por ciertos modos de abusar del lenguaje– reflexionaría luego Wittgenstein en sus *Investigaciones Filosóficas*:

> El sentimiento de la insuperabilidad del abismo entre la conciencia y los procesos en el cerebro, ¿cómo es que esto no juega ningún papel en las consideraciones de la vida corriente? (Wittgenstein 1953: §412)

Es innegable que esa imaginería geográfica del s. XIX –abismos, golfos–, con sus tormentosas evocaciones románticas, tiende hoy a resultar altisonante. Cierta desengañada sobriedad hace que prefiramos, como Levine (1983), hablar de un hiato explicativo –o un salto, una brecha, un intersticio: un *explanatory gap*– entre lo fisiológico y lo mental. Pero no cabe sino dar la razón a Tennant (2007: 754) cuando, después de trazar la genealogía de la metáfora, señala que, tratándose de describir una separación insalvable, hablar de un abismo resulta mucho más preciso que hacerlo de un intersticio.

Sea como sea, lo cierto es que sólo dos años después de que la existencia de tal abismo quedara consagrada en el desalentado dictamen que proclamaría, en la estela de Tyndall y Huxley, Émil du Bois-Reymond –*¡ignorabimus!*–, un joven profesor de filosofía de la Universidad de Würzburgo, Franz Brentano, aprontaba su explicación de la refractariedad de la naturaleza misma de la mente a las indagaciones de la ciencia natural. La clave –como no podía ser de otro modo en quien había comenzado su prometedora carrera académica con una disertación *Sobre los múltiples significados del ente según Aristóteles* (Brentano 1862)– había de buscarse en la tradición escolástica:

> Todo fenómeno mental se caracteriza por lo que los escolásticos de la Edad Media llamaron la in-existencia intencional (o mental) de un objeto, y lo que podríamos llamar, aunque no de manera completamente ambigua, la referencia a un contenido, la dirección hacia un objeto […] o hacia una objetividad inmanente. Todo fenómeno mental incluye algo como objeto dentro de sí mismo, aunque no todos lo incluyen de la misma manera. En la presentación se presenta algo, en el juicio se afirma o se niega algo, en el amor es amado, en el odio odiado, en el deseo deseado y así sucesivamente. Esta in-existencia intencional es característica exclusivamente de los fenómenos mentales. Ningún fenómeno físico exhibe algo parecido a esto. Podemos, por tanto, definir los fenómenos mentales diciendo que son aquellos fenómenos que contienen un objeto intencional dentro de ellos mismos. (Brentano 1874/1973: 88)

La sospecha que encendió en su día Émil du Bois-Reymond, y que las acotaciones de Brentano avivaron, ha impregnado desde entonces, *pace* Cajal, la reflexión filosófica acerca de la próspera tarea investigadora de la psicología. Bien es verdad que sólo tres o cuatro de los siete enigmas en que du Bois-Reymond fundamentara su agnosticismo –los siete *Welträtsel*– atañen a la naturaleza de lo mental. Dos de ellos, además, quedaban planteados bien como un mero problema que el avance del evolucionismo habría de permitir resolver más pronto o más tarde –así, el enigma IV: la aparición de disposiciones teleológicas en los seres vivos–, bien como un misterio subordinado a otro de los siete –todo lo que pudiera haber de inexplicable en relación con la naturaleza del pensamiento y el lenguaje, el enigma VI, se difuminaría si se daba por supuesta una explicación convincente de la naturaleza de las sensaciones, que constituía el enigma V. Quedaban aun así, según el recuento de du Bois, dos misterios que verdaderamente nos trascienden y que atraviesan el ámbito de la mente: la naturaleza de la sensación y la posibilidad del libre albedrío –el séptimo y último enigma de su manifiesto agnóstico.

Pues bien, en torno al primero de estos dos –el enigma V, la naturaleza de la sensación– y en torno a la cuestión del origen del pensamiento y el lenguaje –el enigma VI, que du Bois supeditaba al precedente–, ha orbitado buena parte de la discusión posterior: en particular, en torno al carácter vivencial o experiencial de nuestras impresiones sensoriales, que hemos dado en llamar su *quale* y que parece coincidir con lo que tales impresiones tienen de inefable, con lo que no logramos explicar con palabras acerca de cómo es un cierto dolor, el sabor de una fruta o el enamoramiento, y en torno al rasgo acaso más prodigioso que pensamiento y lenguaje comparten, su capacidad de *referirse* a realidades distintas de la suya propia. *Inteligencia* y experiencia directa serían también para Herbert Feigl (1958: 419) los criterios decisisivos para distinguir lo mental de lo físico –criterios a su entender lógicamente ortogonales, si bien mostrarían de hecho un cierto solapamiento en ámbitos como el de las imágenes mentales–, y ese breve inventario de incógnitas iría asentándose casi como lugar común a la vez que se desplazaba paulatinamente hacia sus dos núcleos supuestos –intencionalidad y consciencia. Lo compendia *inter alia* Clapin (2002):

> Two mysteries continue to frame debate in contemporary philosophy of mind. The first is the nature of consciousness. In particular, how can our conscious experience fit into the growing body of scientific knowledge about the mind and the brain? The second mystery is intentionality. How can our thoughts be *about* other things? (Clapin 2002: 1)

El vasallaje que du Bois conjeturase entre un asunto y otro, sin embargo, parece, de la mano de Feigl, haberse desvanecido. Antes bien, es una suposición muy extendida, al menos en las disquisiciones recientes sobre la naturaleza de lo mental que toman como inspiración los desarrollos de la psicología científica, que la cuestión de la consciencia y la de la intencionalidad son independientes entre sí, o que pueden tomarse como tal a efectos de su esclarecimiento. Como suele ocurrir, han sido quienes pretendían advertirnos contra esa suposición los que más han contribuido a dibujar sus contornos: así, McGinn (1988: 299), que la describe como "the insulation strategy", o Searle (1992: 153), que enmarca su denuncia en una impugnación general de las prácticas explicativas del cognitivismo[5]. La denuncia del arrinconamiento de la investigación sobre la conciencia era patente también en Dennett (1991a) –no es aventurado, de hecho, afirmar que ha acabado por convertirse en un *tópos* de la filosofía de la mente de las últimas décadas. Bajo la mirada de Dennett, ese arrinconamiento aparece vinculado a la tendencia a desarrollar modelos de procesos psicológicos periféricos, postergando en cambio las instancias de procesamiento central. Ese sesgo, en el que incurrirían por igual psicólogos, fisiólogos e investigadores en inteligencia artificial, tiene el desdichado defecto de "[…] leaving too much of the mind's work to be done 'in the center,' […]", lo que a su vez nos conduce a subestimar "[…] the 'amount of understanding' that must be accomplished in the relatively peripheral systems of the brain" (Dennett 1991a: 39)[6].

Desoyendo, en todo caso, esas admoniciones, y aunque sólo sea por cierta modestia epistemológica, es común incluso que se plantee el proyecto de explicar la relación existente entre creencias, deseos, temores o anhelos y lo creído, deseado, temido o anhelado –la relación de intencionalidad– antes de abordar siquiera la pregunta por la consciencia, que aparece más intratable. Así lo plantea, por ejemplo, Pylyshyn (1984): al percatarse de que su interés por el papel que desempeña el concepto de representación en el tipo de explicación abanderado por el cognitivismo lo arroja "[…] right in the middle of what is probably the second hardest puzzle in

[5] Para una valoración tentativa de esa tendencia, y en particular de las advertencias de Searle en su contra, *cf.* Chacón y Hermoso (2009).

[6] Las tesis de Dennett a este respecto se revelan nítidamente deudoras de las conclusiones de Fodor (1983: *passim*; *cf.* por ejemplo 103, *infra*) respecto a la menesterosidad que aqueja a nuestras teorías de los procesos psicológicos centrales en comparación con la relativa riqueza de nuestros modelos de determinados procesos periféricos que, según la hipótesis de Fodor, se caracterizan por su modularidad –encapsulamiento, especificidad de dominio, obligatoriedad. Si bien Dennett se muestra extremadamente crítico con la hipótesis de modularidad –como, en general, con la concepción de lo mental preconizada por Fodor, *cf.* por ejemplo Dennett (1991b, *infra*)–, que los términos en que se plantea su denuncia del arrinconamiento de la conciencia están en buena medida moldeados por dicha hipótesis resulta transparente sobre todo en Dennett (1984b), una reseña de Fodor (1983) que el propio Dennett (1991) cita al hilo de su razonamiento.

philosophy of mind: the puzzle of meaning […]", deja constancia de que, a su entender, "[…t]he hardest puzzle […] is consciousness, which probably is not even well enough defined to qualify as a puzzle" (Pylyshyn 1984: 23). Más irónicamente expeditivo, si cabe, se muestra Fodor (2000), abundando en sus tempranas advertencias acerca de las radicales limitaciones de la concepción computacional de lo mental (*cf.* Fodor 1975, 1983) que él mismo, pese a ello, ha defendido con inigualable vigor:

> It is rather an embarrassment for cognitive science that any intentional mental states are conscious. "Why aren't they all unconscious if so many of them are?" is a question that our cognitive science seems to raise but not to answer. Since, however, I haven't the slightest idea what the right answer is, I propose to ignore it. (Fodor 2000: 106) [7]

Parece, sea como sea, que el diágnostico al uso, cristalizado en Clapin (2002: 1, *supra*), difiere levemente del que más de veinte años antes formulara Field (1978), quien hacía gala acaso de mayor pulcritud a la hora de diferenciar las propiedades relacionales –es decir, *intencionales*: caracterizadas por *tender a* un objeto– que poseen ciertos estados mentales de sus propiedades experienciales o cualitativas: el problema de la consciencia atañe sin duda a ambas, aunque lo haga tal vez –como trataría de argumentar Block (1995)– en sentidos o bajo prismas diferentes. Sea como sea, Field enlazaba sin ambages el problema de la intencionalidad, tal como venía ya entonces aguijando los bueyes de la filosofía de la mente de inspiración analítica, con las preocupaciones de Brentano[8]:

> The unsatisfactoriness of postulating irreducibly mental properties is the source of the two main problems in the philosophy of mind. The first and more widely discussed problem is the problem raised by *experiential* properties, for instance the property of feeling pain: a materialist needs to provide a believable account of such properties according to which those properties are not irreducibly mental. The second problem,

[7] En un tono parecido, mesuradamente jocoso, apunta también Tennant (2007: 746) que a día de hoy "[…p]erhaps the hardest problem in contemporary philosophy of mind, after the problem of consciousness itself, is how to find anything new to say about the problem of consciousness". La exigencia de dar con la respuesta correcta –eso sí– ha quedado ya desbancada por la de dar con una respuesta novedosa, lo que podría tomarse por un signo de estancamiento del debate si no fuera porque parece connatural al carácter histórico del trabajo de Tennant.

[8] En el pensamiento de Brentano pueden encontrarse también las fuentes, desde luego, del acercamiento fenomenológico a la naturaleza de la intencionalidad, que toma como piedra angular la indisolubilidad de su vínculo con la consciencia –no en vano, Edmund Husserl asistió cuando estudiaba en la Universidad de Viena, como es sabido, a algunos cursos dictados por Brentano, y muchas de sus tesis sobre la estructura de la consciencia parecen provenir de aquellos apuntes. La vastísima y ubérrima comarca de reflexión filosófica labrada en los surcos del trabajo de Husserl no será hollada en este estudio, ni siquiera si los razonamientos o las conclusiones ensayadas parecieran concordar con principios medulares de la aproximación fenomenológica, o vulnerarlos. Un minucioso recorrido por las distintas concepciones de la intencionalidad, tanto de inspiración fenomenológica como de corte analítico, que han punteado el pensamiento filosófico del s. XX puede hallarse en Paredes (2007).

> raised by Brentano […], is the problem of *intentionality*. Many mental properties –believing, desiring, and so forth– appear to be *relational* properties: more precisely, they appear to relate people to non-linguistic entities called *propositions*. (Field 1978: 34)

Entre la noción de intencionalidad y la necesidad de dar cuenta del origen y naturaleza del error –de dar cuenta, en definitiva, del hecho de que podamos albergar creencias erróneas acerca del mundo que, no obstante, preserven esa cualidad de ser acerca del mundo– existe un íntimo nexo que habremos de ir perfilando poco a poco. Comoquiera que hay sin duda rasgos que comparten una creencia errónea y una creencia sobre algo que no existe, estos lazos entre la noción de intencionalidad y la naturaleza del error se hallaban ya presentes en la fábrica de la in-existencia intencional que ingeniosamente tejiera de retales escolásticos el propio Brentano. Como ha acertado a resumir Bechtel (1988), resulta evidente que:

> Al señalar el hecho de que los objetos mentales pueden dirigirse hacia objetos o eventos no existentes, Brentano estableció una difícil tarea para los pensadores siguientes. […] Parece que estamos comprometidos con las afirmaciones inconsistentes de que, por una parte, los estados intencionales incluyen una relación con un objeto y de que, por otra parte, el objeto con el que podríamos esperar poner en relación los estados intencionales no necesita existir. (Bechtel 1988: 65)

Someter a examen la capacidad del concepto de representación mental –que ha venido sirviendo de fundamento a la psicología de inspiración cognitivista– para afrontar el reto de Brentano entraña, por tanto, poner de manifiesto las graves dificultades que dicho concepto atraviesa cuando es llamado a sustentar tanto creencias sobre el mundo como creencias sobre entidades inexistentes, o, más aún –aunque Bechtel no siempre diferencie con nitidez ambos casos–, tanto creencias verídicas como creencias falsas sobre cosas mundanas:

> Desde la perspectiva de la ciencia cognitiva moderna, podría suponerse que el problema de la intencionalidad se podría resolver postulando representaciones como los objetos de los estados mentales y, por consiguiente, como los objetos del pensamiento. Aunque las representaciones […] pueden desempeñar un papel importante al explicar cómo es posible la intencionalidad, no pueden desempeñar el papel para el que Brentano parece estar postulando objetos intencionales. La razón puede apreciarse si nos concentramos en las creencias verídicas. En tales casos queremos decir que nuestras creencias son *sobre* el objeto o estado de cosas efectivamente existente en el mundo. Mas, si hacemos de las representaciones los objetos de nuestras creencias en el caso de falsas creencias, un razonamiento similar exige que consideremos las representaciones como los objetos de creencia en el caso de las creencias verídicas. Pero esto no logra capturar el importante elemento para el que se introdujo en primer lugar el término *intencionalidad*, a saber: la idea de que el objeto de nuestros estados mentales son a menudo cosas externas a nosotros. Si adoptamos la herramienta de las representaciones mentales, tenemos aún que explicar cómo algunos de nuestros estados mentales tienen éxito al conectar con las cosas del mundo mientras que otros no logran hacer esto. (Bechtel 1988: 66-67)

Actitudes, proposiciones, hechos

La necesidad de comprender los lazos que traban nuestros pensamientos con las cosas acerca de las que versan, a través acaso de esa mediación abstracta de las proposiciones, o de alguna de semejante índole, configura lo que a veces se ha querido condensar bajo la rúbrica del problema de la mente y el mundo. Si bien se trata –es obvio– de un asunto distinto de aquel, tan antiguo como la propia inquietud filosófica, que gira en torno a la relación entre el alma y el cuerpo, o el espíritu y la materia, o, más recientemente, la mente y el cerebro, no es menos obvio que los pasadizos entre ambas cuestiones son abundantes. En su destilada versión contemporánea, despojado de las zozobras acerca de la inmortalidad del alma que en su origen lo animaban, o de las dudas en torno a si la admirable concordancia entre sus peripecias y las del cuerpo se debía a la constante intervención divina que sospechaba Malebranche o a la perfectísima armonía que, en la imaginación prodigiosa de Leibniz, punteaba desde siempre cuanto existe, el problema mente-cuerpo se articula fundamentalmente –como ha dado en resumir Rabossi (1995)– en tres enigmas: "la *naturaleza* de los fenómenos mentales *vis-à-vis* la naturaleza de los fenómenos físicos […]; […] el *status ontológico* de los fenómenos mentales […], y […] la eventual *relación* de los fenómenos mentales con los fenómenos físicos" (Rabossi 1995: 17). Un cabal entendimiento de las condiciones que han de cumplirse para que pueda afirmarse con verdad que tal o tal otro estado de cosas –existente o no– constituye el objeto de una sensación, de una creencia, de un deseo o de una intención, sin duda iluminaría, en más de uno de los sentidos enumerados por Rabossi, nuestra comprensión del lugar de lo mental en la urdimbre de la realidad. Así, según venimos de repasar, lo intuía desde luego Brentano (1874): su notoria respuesta a la pregunta por la naturaleza de los fenómenos mentales, por aquello que los distingue esencialmente de los puramente físicos, es que aquéllos, a diferencia de éstos, son *acerca de otra cosa*: "incluye[n] algo como objeto dentro de sí" (Brentano 1874, *supra*).

Con el análisis de Chisholm (1957), proseguido con distintos matices en Chisholm (1958, 1967, 1980, 1991), cobra fuerza como es bien sabido un giro desde el estudio de los actos mentales que habían interesado a Brentano –o los estados mentales que favorecería la tradición estructuralista– hacia el lenguaje que usamos para referirnos a algunos de esos fenómenos psicológicos. Es ya materia de cualquier curso elemental de filosofía de la mente la observación de que al atribuir a alguien una creencia o un deseo, la verdad de la atribución no depende de la verdad de la creencia atribuida, o de la satisfacción del deseo; ni siquiera de la existencia de aquello acerca de lo que se cree algo, o de aquello que se desea. También la noción de opacidad referencial se ha agregado al *corpus* de conocimientos básicos provenientes del análisis del lenguaje psicológico cotidiano: cuando decimos de alguien que cree o desea tal o cual cosa, no cabe esperar que la atribución conserve intacto su valor de verdad aunque reemplacemos la descripción de lo creído o deseado por otras que se

refieran de hecho exactamente a lo mismo, porque esa identidad de referencia bien puede ser opaca para el sujeto que alberga –o al que le atribuimos– la creencia o el deseo. Así que podemos, con verdad, decir de Edipo que desea casarse con Yocasta pero también que no desea casarse con su madre, por mucho que Yocasta y la madre de Edipo sean la misma persona, porque y en la medida en que esa correferencialidad es ignorada por el infortunado rey de Tebas. Es habitual que esta misma noción de opacidad referencial quede descrita, en otros términos, como la testaruda insubordinación de los enunciados intensionales al principio de sustituibilidad *salva veritate* que conocemos como ley de Leibniz, y que rige sobre aquellos cuyo contenido resulta ser estrictamente extensional: *eadem sunt, quorum unum potest substitui alteri salva veritate* –el principio que, según aprendimos en la escuela, nos permite resolver sistemas de ecuaciones sustituyendo una expresión por otra que de acuerdo con la ecuación tiene idéntico valor: despejar y sustituir, nos decían. Pues bien, estas peculiaridades lógicas delatarían, a juicio de Chisholm, la presencia de fenómenos psicológicos en las oraciones que las exhiben[9]:

> […] when we wish to describe certain psychological phenomena –in particular, when we wish to describe thinking, perceiving, believing, seeing, knowing, wanting, hoping, and the like–, either (a) we must use language which is intentional or (b) we must use a vocabulary which we do not need to use when we describe non-psychological, or "physical", phenomena (Chisholm 1956: 129; *cf*. Chisholm 1957: 172-173 y Chisholm 1991: 298b)

Cabe, en efecto, pensar en las creencias, los deseos, las esperanzas, los temores o las certezas como distintos tipos de *actitudes* que podemos entablar hacia diversas *proposiciones* –adoptamos, por ejemplo, la actitud de *creer* hacia la proposición "Todos los hombres son mortales", como hacia tantas otras; bien podemos adoptar también actitudes distintas hacia esa misma proposición. Pensar así acerca de nuestros estados mentales nos conduce a la idea, introducida por Russell (1940), de que al menos una buena parte de ellos son *actitudes proposicionales*. Dicha idea es reconocidamente ubicua en el cognitivismo:

> Philosophers and cognitive scientists often use the term *propositional attitude* as a general label for those mental states that have conditions of satisfactions –states like believing, desiring, wishing and fearing. According to the most widely held theory in this area, what it is for a person to have a propositional attitude is for the person to stand in an appropriate relation to a special kind of internal state –a *mental representation*. (Warfield y Stich 1994b: 3)

[9] Aunque, desde luego, no sólo el lenguaje psicológico genera contextos intensionales: lo mismo puede ocurrir con los términos que aparecen mencionados en el discurso en lugar de usados –o sea, propiamente entrecomillados–, o con los que se encuentran bajo el alcance de operadores lógicos de modalidad, como "posible" o "necesario". Ya Frege (1892, *infra*) consideraba el lenguaje psicológico sólo como un caso particular de *oratio obliqua*.

Desde luego, es aquello que creemos –deseamos, tememos…– lo que utilizamos para describir nuestras creencias –deseos, temores…– en la vida cotidiana, así como para decidir si nos referimos a la misma o a distintas creencias. Aunque –como ha señalado acertadamente Dennett (1982)– un uso sofisticado del lenguaje permita otros modos de mencionar nuestros propios estados mentales, lo cierto es que:

> […] the privileged way of referring to beliefs, what we usually mean and are taken to mean in the absence of special provisos or contextual cues, is *the proposition believed*: e.g., the belief that snow is white, which is *the same belief* when believed by Tom, Dick, and Harry, and also when believed by monolingual Frenchmen –though the particular belief-*tokens* in Tom, Dick, Harry, Alphonse and the rest […] might differ in all sorts of ways that were not of interest to us given our normal purposes in talking about beliefs. (Dennett 1982: 3)

Ahora bien, parece obvio, al menos a primera vista, que la jerga felizmente acuñada por Russell no debe entenderse como una *explicación* de la intencionalidad. Al igual que ocurre con el concepto de representación, el de proposición pertenece más al ámbito de la topografía del *explanandum* que al del desarrollo del *explanans* –o, al menos, su desempeño explicativo no se acredita en la mera definición del concepto. Con palabras prestadas de Rodríguez (2001b: 225): "[…] hasta aquí nada más que hemos constatado hechos, otra cosa es hacerlos inteligibles". Ciertamente,

> […] el trabajo crítico de explicar la intencionalidad no se hace postulando la proposición o la representación. La tarea de explicar cómo las proposiciones o las representaciones son *sobre* objetos o eventos del mundo, algunos de los cuales no existen de modo efectivo, queda por realizar. (Bechtel 1988: 73)

También Blanco (2001: 108) ha recalcado la misma distinción:

> […] una cosa es *presuponer* la intencionalidad de los estados mentales que intervienen en los procesos mentales, y otra muy distinta es *explicar* de dónde viene esa intencionalidad.

Por otro lado, ya Quine (1960: 176, 199-201) advirtió que las irregularidades lógicas en que incurren los verbos que designan actitudes proposicionales se dan también en verbos como "cazar", que no designan estados mentales sino conductas: por eso podría tener sentido la afirmación de que alguien está cazando unicornios. También "mirar", como nos enseñaron unos versos de Bertold Brecht, se comporta de ese modo[10]. No es menos cierto, sin embargo, que conceptualizar nuestros deseos y creencias, nuestras esperanzas y temores, como actitudes proposicionales –acaso

[10] "Recuerdo de María A.", en *Libro de plegarias domésticas* (Brecht, 1927):

> Fue un día del azul septiembre cuando / bajo la sombra de un ciruelo joven, / tuve a mi pálido amor entre los brazos / como se tiene a un sueño calmo y dulce. / Y en el hermoso cielo de verano, / sobre nosotros, contemplé una nube. / Era una nube altísima, muy blanca. / Cuando volví a mirarla, ya no estaba. […].

también entender bajo ese prisma algunas de nuestras acciones– nos instiga a escrutarlas, o a inspeccionar al menos su componente proposicional, desde la perspectiva del análisis lógico del lenguaje. Dicho análisis lingüístico se ha erigido, sobre todo –según se ha dicho– a partir de los trabajos de Chisholm (1957), en una copiosa fuente de averiguaciones acerca de la naturaleza de la intencionalidad. En suma,

> […] la armazón de las actitudes proposicionales sugiere también un modo de caracterizar la intencionalidad de los estados mentales: usamos la proposición hacia la que la persona tiene una actitud para identificar el contenido del estado mental de la persona. El uso de proposiciones para especificar el contenido de los estados mentales sugiere una conexión entre los análisis del lenguaje y de la mente. (Bechtel 1988: 71)

Bien puede de hecho entenderse, tal como ha sugerido entre otros Bechtel (1988: 73, 78), el desarrollo del cognitivismo que pareció, durante cierto tiempo, erigirse como canónico –la teoría computacional de la mente, ligada a la hipótesis del lenguaje del pensamiento (Harman 1970, Fodor 1975, Field 1978)– precisamente como un intento de atesorar los frutos explicativos de la homología entre explicación psicológica y análisis lingüístico a la que apunta la noción de actitud proposicional. Pero el "modo de caracterizar la intencionalidad de los estados mentales" que insinúa el esquema de Russell está lejos de ser, desde luego, una primicia que se desprenda de él: no es, como ha señalado Dennett (1982: 3, *supra*) más que la táctica que cotidianamente empleamos para referirnos a un buen número de estados mentales. De ahí, naturalmente, que Fodor haya presentado reiteradamente su proyecto –al menos desde Fodor (1968)–, y en especial la propia hipótesis del lenguaje del pensamiento, como una vindicación de la psicología del sentido común, o que Pylyshyn (1984: 2589 lo haga aduciendo que "[…] buena parte de la psicología de las abuelas es sólida, si no en su explicación de los mecanismos subyacentes, sí en los *tipos* de regularidades a las que apela"[11].

Así las cosas, parece formar parte de las ordenanzas de la tarea de dar fundamento ontológico y epistemológico a las explicaciones psicológicas desplegadas en el seno del cognitivismo la asunción de que la idiosincrasia del lenguaje que usamos coloquialmente para hablar de nuestras creencias y deseos ha de contemplarse bajo el prisma de la intensionalidad de sus expresiones, que contrastaría con la extensionalidad de las empleadas en otros usos del lenguaje. El

[11] O también que, en sentido inverso, Kintsch, Miller y Polson (1984) hayan podido ver en esta reivindicación –tal como nos recuerda Rivière (1991b: 144)– la "[…] equivocidad constitutiva" del cognitivismo, que habría amasado buena parte de su pujanza "[…] al mezclar los conceptos propositivos y mentalistas de la psicología natural […] con las categorías derivadas de la noción de cómputo". Parece coincidir con ellos McDowell (1994b), quien ha argumentado que el intento de elucidar nociones de la psicología tradicional mediante las de la psicología cognitiva, o viceversa, sólo lleva a la confusión. Hasta qué punto la divergencia entre el aparato conceptual de la psicología ordinaria y el de la psicología cognitiva resulte asimilable para distintas interpretaciones metateóricas de esta última es una cuestión sobre la que será preciso regresar *infra*.

cotejo de la intensión y la extensión de un término o un enunciado lingüístico arraiga –es bien sabido–, en la práctica de la *definitio fit per genus proximum et differentiam specificam* a la que Aristóteles diera carta de naturaleza en sus *Tratados de Lógica* u *Organon* y cuyo propósito es la sucesiva delimitación de la intensión de un concepto –las *notas* que lo definen, se diría– pero no la enumeración exhaustiva de su extensión. *Hombre* es, entonces, el animal, bípedo, implume *y de uñas planas* con que Platón devolviera al Cínico la burla de dejarse ver por la Academia con un gallo desplumado a cuestas y diciendo "Éste es el hombre de Platón" (Diógenes Laercio, *Vidas* VI: 14), tanto como *hombres* son Sócrates, Diógenes y un extenso etcétera que alcanzase hasta el primer hombre, fuera éste Foroneo, como contaba Solón en *Timeo* 22*a*, o Pelasgo, como relataría Pausanias (*Descripción* VIII: 1.2). La distinción entre, por así decirlo, delimitar un conjunto acotando los rasgos compartidos sólo por los elementos que incluye y hacerlo enumerando esos elementos quedaría reescrita con caracteres modernos al diferenciar Mill (1843) en cada nombre su connotación –de la que, a su entender, carecerían los nombres propios– y su denotación; de ahí pasaría a los manuales de lingüística, a veces con el matiz de denominar denotación a la relación entre una expresión y su extensión y connotación a la que se da entre la expresión y su intensión. Bajo la ancha sombra de Frege, por fin, la distinción entre intensión y extensión se enlaza con la que media entre el sentido y la referencia de un signo, introducida en su trabajo sobre "Función y concepto" (Frege 1891) y detallada en un ensayo que redactó a la par que el primero, "Sentido y referencia" (Frege 1892). Con ello, la noción de intensionalidad cobra de pronto una plétora de nuevos matices, derivados de la complejidad y delicadeza de la conceptografía fregeana y de sus reflexiones sobre semántica y filosofía de la lógica, que paulatinamente irán tiñendo la investigación sobre la naturaleza de lo mental a la luz de las peculiaridades del lenguaje que lo describe.

La sombra de Frege

La idea de sentido de un signo nace en Frege del esfuerzo por dilucidar la naturaleza de la relación de identidad que en una ecuación se expresa con "=". La poderosa intuición que Frege esgrime es que el hecho de que ese símbolo sea siquiera necesario se debe enteramente a que los que lo flanquean –los términos de la ecuación– pueden referirse exactamente a un mismo contenido pero determinándolo de maneras dispares. Si hubiera sólo un modo en que un contenido cualquiera pudiera venir determinado por un símbolo, entonces ningún enunciado de identidad podría ser verdadero más que enlazando instancias del mismo símbolo –como en "*a*=*a*"–, y dichos enunciados resultarían siempre superfluos. Pero dado que podemos referirnos a –digamos– Aristóteles como "el hijo de Nicómaco y Efestiada que estudió con Platón" con tanta verdad como con "el autor de *De Anima*", necesitamos un símbolo, "=", que nos permita postular la relación existente entre los múltiples signos que pueden, de distintas formas, coincidir en una misma referencia –el propio

Aristóteles, en el ejemplo. A lo que en el orden semántico no comparten dos signos distintos que sí comparten su referencia (su *Bedeutung*) es a lo que Frege llamará el sentido (*Sinn*) de cada uno de ellos. El sentido es, pues, un modo de determinar la referencia del signo, ya sea mencionando, en el caso de Aristóteles, lo sustancial de su genealogía, ya alguno de los escritos que nos legó y damos como suyo. "Los signos" –nos recuerda Kenny (1995: 167)– "[…] expresan sus sentidos y denotan sus referencias. Al usar signos expresamos un sentido y denotamos una referencia".

Provisto de estas herramientas, Frege no tardó en ocuparse del extravagante comportamiento de las oraciones que mencionan creencias o deseos, que después preocuparía a Chisholm (1957, *supra*). Esas irregularidades eran para Frege (1892) parte de un campo más amplio, el de la *oratio obliqua*, en el que se incardinaban la cita directa, la cita indirecta y las expresiones que Russell bautizaría como descripciones definidas, del estilo de "el hijo de Nicómaco y Efestiada". Las expresiones en las que se articulan todos estos giros –argumentaba Frege– pierden su referencia habitual, que de acuerdo con el análisis laboriosamente desplegado por el propio Frege era un valor de verdad (*cf. infra*), y, con ello, dejan de ser susceptibles de la interpretación veritativo-funcional canónica. "Edipo desea casarse con Yocasta", o "Duncan creía que Macbeth era digno de confianza" (Kenny 1995: 179), se revelan en este contexto como casos de cita indirecta en los que –igual que en "Tales dijo que el agua es el principio de todas las cosas"– la cláusula que describe el objeto del deseo, la creencia o la declaración no se refiere a su propio valor de verdad sino a su sentido. De ahí, entonces, que un principio veritativo-funcional como es la ley de Leibniz no rija para expresiones de esta índole, o que éstas carezcan de compromiso con la verdad de la expresión que las dota de intensionalidad, o con la existencia de sus referentes. De ahí, en suma, que podamos atestiguar que, como deja casi esquematizado Kenny (1995):

> Las proposiciones "Copérnico creía que las órbitas de los planetas eran circulares" y "Copérnico creía que la Tierra gira alrededor del Sol" son ambas verdaderas, pese a que el contenido de la primera cláusula "que" [completiva] es falso y el de la segunda verdadero. Por otra parte, aunque la proposición "Urano gira alrededor del Sol" es tan verdadera como "Venus gira alrededor del Sol", la proposición "Copérnico creía que Venus gira alrededor del Sol" es verdadera, mientras que la proposición "Copérnico creía que Urano gira alrededor del Sol" es falsa, puesto que Urano no había sido descubierto en los días de Copérnico. (Kenny 1995: 179)

O bien, con las palabras –igual de sintéticas, mucho más abstractas– empleadas por el propio Frege:

> La oración principal junto con la subordinada sólo tiene como sentido un único pensamiento, y la verdad del todo no incluye ni la verdad ni la no verdad de la oración subordinada. En estos casos, no se permite reemplazar en la oración subordinada una expresión por otra que tenga la misma referencia habitual, sino sólo por una que tenga la misma referencia indirecta, es decir: el mismo sentido habitual. (Frege 1892: 96)

De la mano de Frege –y, por tanto, de un decidido antipsicologismo– la idea de intensión habría de enlazarse también con las de pensamiento e imagen mental. Cuando entendemos una palabra, o una expresión, lo que hacemos es captar su sentido; habiéndolo captado, nos es dado entonces determinar su referencia, si la tiene –hay, por supuesto, sentidos sin referencia, como el de la palabra "Pegaso" o el de "el mayor número primo". Pero el sentido no es una imagen mental –la que cada uno de nosotros, pongamos por caso, pueda tener de Pegaso– puesto que éstas, a diferencia de aquel, son mudables e idiosincrásicas. De acuerdo con la elegante metáfora dibujada por Frege:

> Alguien observa la Luna a través de un telescopio. Comparo la Luna misma con la referencia; es el objeto de observación, que viene dado por la imagen real que se proyecta en la lente del objetivo del interior del telescopio y por la imagen retiniana del observador. A la primera imagen la comparo con el sentido; a la segunda, con la representación o intuición. La imagen del telescopio es […] objetiva […]. […] Pero, por lo que respecta a las imágenes retinianas, cada uno tendría la suya propia. (Frege 1892: 89)

Si no una imagen, el sentido ha de ser un *pensamiento*; el sentido, en particular, de una aseveración o proposición asertórica completa, de un juicio, es un pensamiento –no así el de otras expresiones lingüísticas, como los ruegos o las promesas. Pero el pensamiento no puede venir constituido sólo por la referencia de tales expresiones, como muestra el hecho de que "Aristóteles fue el hijo de Nicómaco y Efestiada que estudió con Platón" y "Aristóteles fue el autor de *De Anima*" son a todas luces distintos pensamientos –pues uno podría, por ejemplo, tomar por verdadero el segundo y por falso el primero, mientras que no podría hacer tal cosa si se tratara de un solo pensamiento–, y, sin embargo, "Aristóteles fue el hijo de Nicómaco y Efestiada que estudió con Platón" y "Aristóteles fue el autor de *De Anima*" no pueden tener distintas referencias, ya que sólo difieren en la descripción definida que completa el sintagma verbal, y ambas descripciones definidas, teniendo como tienen la misma referencia, han de contribuir de igual manera a la referencia global de la expresión. En esta encrucijada se dirimen dos de las conclusiones más arriesgadas –y también más relevantes– del análisis de Frege: si el pensamiento expresado por un juicio no puede ser su referencia, habrá de ser su sentido; si la referencia de un juicio no puede ser el pensamiento expresado, entonces ha de ser su valor de verdad. De igual modo que la referencia de "El hijo de Nicómaco y Efestiada que estudió con Platón" es: Aristóteles, la referencia de "Aristóteles fue el hijo de Nicómaco y Efestiada que estudió con Platón" sería: verdadero.

Diremos, entonces, que el sentido de una expresión determina su extensión, que porta un valor de verdad, que constituye el pensamiento que la mente aprehende al entender la expresión. La titánica carga confiada a los hombros de la noción de sentido provocaría importantes dificultades, que el propio Frege en buena parte detectó y que, en ocasiones, siguen, como se verá, moldeando el debate en

torno a la intencionalidad de lo mental a día de hoy[12]. Una de esas dificultades asoma, precisamente, en el contexto de un argumento que Frege aduce para defender su controvertida tesis de que la referencia de una proposición asertórica completa es su valor de verdad. Cabría pensar, como hipótesis alternativa, en que tales proposiciones tengan sólo sentido, y no referencia, al igual que ocurre con términos como "Pegaso". Pero, piensa Frege, dado que tal cosa parece verdadera de algunas afirmaciones, como "Pegaso vuela", y que nos es preciso diferenciar esa clase de afirmaciones de otras como "Los caballos vuelan" o "Los caballos no vuelan", no podemos aceptar que todas ellas carezcan de referencia: habremos de concluir más bien que "Pegaso vuela" carece de referencia, pero que "Los caballos vuelan" tiene como referencia: falso, y "Los caballos no vuelan" tiene como referencia: verdadero. La ausencia de referencia de "Pegaso vuela" puede trazarse, obviamente, a la de "Pegaso" –como, en el ejemplo elegido por Frege (1892: 91), la de "Ulises fue dejado en Ítaca profundamente dormido" puede rastrearse hasta la de "Ulises". Ahora bien, si, como parece, el efecto de la falta de referencia de uno de los términos de una proposición asertórica completa es dejar privado de referencia al enunciado entero, se sigue que tales enunciados tendrán idéntico sentido, expresarán el mismo pensamiento, tanto si todos sus términos son referenciales como si no es así. El pensamiento aprehendido por la mente al entender un enunciado está constituido por el sentido del enunciado, luego la mente aprehende el mismo pensamiento cuando comprendemos la oración "Ulises fue dejado en Ítaca profundamente dormido" –o "Pegaso vuela"– tanto si existió Ulises –o Pegaso– como si no, y, por supuesto, tanto si fuera cierto como si no que los marinos feacios abandonaron a Odiseo, dormido, en un playa de Ítaca o que el caballo amado por las Musas del Helicón podía hollar el cielo con sus cascos en forma de luna. Como apunta Kenny (1995) en su estudio de Frege (1892):

> Si estuviésemos interesados sólo en el pensamiento, no nos importaría que "Ulises" tuviera o no una referencia, pues el pensamiento sigue siendo el mismo en uno y otro caso, ya que está determinado por los sentidos, y no por las referencias, de las partes constituyentes de la proposición.
>
> Sólo cuando se desea seriamente tomar al enunciado por verdadero o por falso es cuando se siente la necesidad de ascribir una referencia a "Ulises"; porque de otro modo no habría nada que nos permitiera decidir si el predicado a él atribuido era verdadero o falso. (Kenny 1995: 170)

O, con Frege, "[…] queremos que todo nombre propio tenga no sólo un sentido sino también una referencia […p]orque, y en la medida que, nos interesa su valor de verdad" (Frege 1892: 92).

De esta manera, la cuestión de la verdad o falsedad de un pensamiento ha quedado desterrada del ámbito de la psicología, como se hará patente a partir de los trabajos de Putnam (1975a) y Fodor (1980a) –*cf. infra*. El propio Frege ahondaba ya en

[12] *Cf.* por ejemplo, Putnam (1975a: *passim*), Dennett (1982: 11), *infra*.

la cuestión en un trabajo de 1918, "El pensamiento", destinado a formar parte de unas *Investigaciones Lógicas* que no llegó a dar por terminadas. De la verdad y de sus leyes –decía Frege– se ocupa la lógica: descubre las formas de inferencia válida, sienta los fundamentos de su reglamentación prescriptiva y "[…] quizá también un elemento en la explicación causal de los procesos mentales reales" (Kenny 1995: 231). En cambio, de las leyes que describen el encadenamiento, en esas secuencias de causa y efecto, de estados y procesos mentales se ocupa la psicología, pero esas leyes no hacen "[…] ninguna distinción entre pensamientos verdaderos y pensamientos falsos" –o, cabe añadir, pensamientos que no determinan referencia veritativa alguna– porque "[…] el error y la superstición tienen sus causas al igual que la[s] tiene el conocimiento correcto" (Kenny 1995: 230).

Hay, sin embargo, una grieta en el razonamiento que nos ha traído hasta aquí. En la metáfora del telescopio y la Luna –recordemos– Frege asimilaba el sentido de una expresión con la imagen de la Luna en el cristal del objetivo; la propia Luna se hacía corresponder con la referencia, y su imagen retiniana con la intuición, la representación mental de la expresión. Pero ahora sabemos que hay pensamientos falsos –sentidos que determinan la referencia veritativa: falso. Sabemos también que la psicología no debe distinguirlos de los verdaderos –al menos para algunos de sus fines explicativos, ya que los efectos de un pensamiento sobre la conducta de quien lo alberga se nos antojan los mismos sea verdadero o falso el pensamiento en cuestión–, pero que la lógica precisa de ellos aunque sea sólo para negarlos, pues –en un hermoso giro parmenídeo del trabajo de Frege[13]– "[…] no puedo negar lo que no es" (Frege 1919: 232). En cambio, la imagen de la Luna en el cristal del objetivo no puede como tal ser *falsa* salvo, precisamente, en la medida en que sea o en que la entendamos como una representación. Es decir: para que esa imagen de la Luna pueda ser falsa debe ser tomada como como *imagen de* la Luna. Entretanto, simplemente, *está ahí* como efecto de las que quiera que sean sus causas: si, tomada como imagen, es verdadera, la luz solar reflejada por la Luna, su trayecto sin perturbaciones por la atmósfera terrestre, el correcto funcionamiento del telescopio…, o cualquier desviación de ese itinerario si la imagen, tomada como tal, es en uno u otro sentido falsa. La imagen de la luna en el telescopio, en fin, no puede ser falsa igual que no puede ser falsa la fiebre salvo que se tome como signo[14], ni pueden ser falsas las anillas del tronco de un árbol salvo bajo una interpretación

[13] *Cf*. por ejemplo, D-K 28 B2, D-K 28 B6. Parece obvio que de la confluencia de sus indagaciones lógicas con las de Frege nacería la interpretación de Parménides que en sus ensayos divulgativos de Historia de la Filosofía presentaría Bertrand Russell (1945), y en torno a la cual se levanta, a partir sobre todo de Owen (1960), la influyente reconstrucción de los fragmentos de Parménides como fruto de preocupaciones de índole más lógica que física, interpretación que matiza toda la lectura de los eleáticos en Barnes (1979, 1982) y en Kirk, Raven y Schofield (1983).

[14] Es decir: si, como es acostumbrado, entendemos como signos las manifestaciones objetivas de una enfermedad y como síntomas aquellos fenómenos que el paciente, en su relato subjetivo del malestar que lo aflige, relaciona con éste, podríamos descubrir que la fiebre de un paciente no viene ocasionada por la enfermedade la que la considerábamos signo, pero no por ello, claro está, dejaría el paciente de tener fiebre ni, en puridad, dejaría la fiebre de ser un síntoma.

dendrocronológica, como índice de su edad. Nos cuesta ver que es así en el caso de la Luna sólo porque se trata de una imagen *en un telescopio*, un instrumento que hemos construido precisamente para obtener *imágenes* astronómicas fidedignas. Se hace difícil entender, entonces, que puedan existir pensamientos falsos, cuando parece atribuírseles, con Frege, una naturaleza objetiva aparentemente irreconciliable con tal falsedad.

En el mismo pasaje de "Sentido y Referencia" donde engarza la metáfora de la Luna y el telescopio queda bosquejada la solución que Frege daría a esta dificultad en sus últimos escritos, más de treinta años después. "La imagen del telescopio" –dice allí Frege– "es […] objetiva en la medida en que puede servir a muchos observadores. En cualquier caso, podría disponerse de tal manera que muchos la usaran al mismo tiempo" (Frege 1892: 89). No reside la objetividad de esa imagen –o del pensamiento que es su tenor en la metáfora– en su verdad, sino en su intersubjetividad, "[…] en la posibilidad de que diferentes pensadores lo capten como uno y el mismo pensamiento" (Kenny 1995: 244).

Bien. Pero lo que parece eludir al penetrante raciocinio de Frege es que la aprehensión del pensamiento –sea verdadero, falso o carezca de referencia veritativa– debe proceder por la vía de la intuición o la representación mental, igual que, en la metáfora, la percepción de la imagen de la Luna en el telescopio discurre por su imagen retiniana. Ahora bien, la imagen retiniana es tan capaz o incapaz de ser *falsa* como la imagen en el cristal del objetivo, la fiebre o las anillas de crecimiento que oculta la corteza de muchos árboles. Naturalmente, lo mismo se predica de los eslabones causales que puedan ir trabándose a la imagen retiniana a través del nervio óptico y el núcleo geniculado lateral, hacia el córtex. Compete pues a la psicología, como uno de sus trabajos fundacionales, explicar *la posibilidad de que aprehendamos pensamientos falsos*, de que tengamos intuiciones, representaciones mentales o, sencillamente, percepciones falsas *como resultado* de ese encadenamiento de causas y efectos en el que verdad y falsedad no son nociones pertinentes. Así que a la psicología, después de todo, al menos desde la perspectiva de su propia fundamentación, sí le concierne la cuestión del error, por mucho que en cuanto atañe a la explicación de la conducta la diferencia entre un pensamiento verdadero y uno falso pudiera resultar inerte.

Por lo demás, el intento de Frege de preservar la noción de pensamiento de todo tinte psicológico ha tenido poco predicamento, al menos –tal vez fuera previsible– en el ámbito de la reflexión sobre la naturaleza de lo mental y las peculiaridades de la explicación psicológica. Indicio claro de ello es que la idea de que el sentido consiste en un modo de *determinar* la referencia, que atraviesa buena parte de los ejemplos didácticos elaborados por Frege, ha quedado a menudo reemplazada por la de que el sentido es el modo de *presentación* de la referencia. Es patente que la idea de presentación de la referencia comporta acentos psicológicos de los que carece la de determinación de la referencia, pues a una presentación parece serle consustancial la presencia de un sujeto que la contemple. No en vano, "presentación" ha sido también en ocasiones la traducción elegida para el término

"Vorstellung", con que Frege se refería a la imagen, intuición o representación interna –el tenor de la imagen retiniana de la Luna en la metáfora del telescopio–; la noción de "modo de presentación", pues, tiene a entremezclar las nociones fregeanas de *Sinn* y de *Vorstellung*. Al aparecer así el sentido como un asunto de trasfondo psicológico, se estrechan los lazos entre el lenguaje que usamos para atribuir actitudes proposicionales, con su característica intensionalidad, y los propios estados mentales atribuidos, de manera que el movimiento que impulsara Chisholm (1957) se ve tonificado.

Sin embargo, algunos perfiles concretos de ese giro hacia el lenguaje psicológico alentado por Chisholm (1957) –así, particularmente, su estrecha dependencia del análisis de una variedad de oraciones psicológicas que, en la estela de la noción russelliana de actitud proposicional, se decreta canónica[15]– han quedado descritos con acento muy crítico por Place (1999), que los considera frutos de un error cometido por Ryle (1949) en su influyente estudio sobre *El concepto de lo mental*. Una observación de Wittgenstein en *Los cuadernos azul y marrón* de sus apuntes de clase de 1933 a 1935, que se publicarían como Wittgenstein (1958), acerca del carácter práctico del conocimiento involucrado en el hecho de saber (cómo) seguir silbando una cierta melodía a partir de una interrupción proporcionó seguramente a Ryle –como señala Place (1999: 373)– la inspiración para diferenciar la noción de "saber cómo" de la de "saber que". Esta distinción –cuyos orígenes pueden trazarse hasta la máxima aristótelica sobre el aprendizaje de las artes y la virtud según la cual "[…] lo que hay que hacer después de haber aprendido, lo aprendemos haciéndolo" (*Ética Nicomáquea* II: 1103*a*)– ha impregnado desde entonces tanto el análisis del lenguaje psicológico como la propia psicología experimental, cuya diferenciación entre conocimiento procedimental y declarativo –prefigurada por Fitts y Posner (1967), ultimada por Anderson (1982, 1983)– es un nítido trasunto de la que perfiló Ryle. Pero, según el diagnóstico de Place, al fijar el contraste entre "saber que" y "saber cómo":

> […] Ryle fails to notice that "knowing how" is just *one* among a number of locutions in which the verb "to know," together with other verbs of cognitive success such as "to remember," takes as its grammatical object an embedded sentence in *oratio obliqua*, or indirect reported speech, which, unlike the case of "knowing that," in which the embedded sentence is declarative, is in the interrogative mood introduced by an interrogative pronoun of which "how" is only one. Thus besides "knowing that" and "knowing how" one can be said to know what, when, where, whether, which, who, and why. (Place 1999: 373)

[15] Se trata, en concreto, de oraciones como "[Yo] creo que va a llover", formadas por un sujeto y un verbo en voz activa que designa un estado de creencia, deseo o afines, y que viene unido por la conjunción completiva "que" a una oración subordinada sustantiva de complemento directo. Más precisamente, lo que queda velado en la idea de estados "afines" al de creencia o deseo parece ser el requisito de que el fenómeno psicológico en cuestión presente una moderada carga epistemológica, a diferencia del que expresarían verbos como "saber", y también, aunque acaso con menor grado de exigencia, una moderada carga afectiva, a diferencia del que se pone por obra en verbos como "lamentar" o "temer". En cualquier caso, el componente epistemológico o afectivo tiende a considerarse ajeno al análisis del estado psicológico en cuestión *en tanto que estado intencional*.

Lo que se avista en el análisis de Place es, por tanto, la diferencia entre mencionar la respuesta a una pregunta que uno conoce –como en "Hamlet *sabía que* Claudio había matado a su padre" y mencionar la pregunta cuya respuesta uno conoce –como en "Hamlet *sabía quién* había matado a su padre"–, pero no la diferencia entre dos formas de conocimiento que parecía desprenderse del planteamiento de Ryle. La única diferencia en el caso de "saber cómo" es que no se espera que la respuesta se efectúe "[…] by giving a verbal account of the procedure to be followed, but by giving a demonstration" (Place 1999: 373). Pero incluso ese último requerimiento podría ponerse en suspenso a la larga para admitir así la atribución "*x* sabe cómo hacer *A*", en ausencia de demostración, como denotativa de la destreza de la que se trate.

Pues bien, la torpeza del análisis ryleano de las peculiaridades gramaticales del lenguaje psicológico –que afectaba también a muchas de las observaciones de Wittgenstein– habría dejado según Place (1999: 374) la puerta abierta para que Chisholm designara las oraciones que describen actitudes proposicionales como linaje privilegiado del lenguaje psicológico, en tanto que puede darse en ellas la in-existencia intencional que Brentano (1874) tomara por marca de lo mental. Entretanto, las consideraciones de Frege (1892) respecto a la *oratio obliqua*, en las que el lenguaje psicológico quedaba circunscrito como un caso particular de cita indirecta, perderían buena parte de su vigor. Así, mientras "saber que" habría quedado relegada como una locución no estrictamente psicológica en virtud de su componente de logro epistémico, giros como "creer que", "desear que" o "temer que" habrían hecho orbitar sobre sí la reflexión en torno a la naturaleza de los estados psicológicos –pues podemos tener creencias falsas, deseos insatisfechos, temores incumplidos, creencias, deseos o temores sobre cosas inexistentes, etc.–, desterrando con ello a otros legítimos objetos de esas reflexiones, descuadernándolas del claro y firme discernimiento de que los contextos intensionales no son privativos del lenguaje psicológico, y por si fuera poco –en una lectura que desborda sin duda las palabras de Place– infectándolas de una distinción de espuria rigidez entre saberes declarativos y procedimentales.

La cuestión del naturalismo

Contra el empeño de Brentano (1874, *supra*), la idea de un *proyecto de naturalización*, heredera del que Quine (1969) vislumbró para la epistemología, había de ir poco a poco enseñoreándose de la semántica, ya tomara ésta como objeto a las palabras o a los pensamientos. El naturalismo, desde entonces, es a menudo asumido como una suerte de canon de perfección epistemológica que serviría de requisito ineludible para dar por buena una teoría de la intencionalidad –como de cualquier otra cosa. Naturalismo es, bajo este prisma, no desistir precozmente del esfuerzo de entender la naturaleza de las cosas –no desistir antes de haberlo llevado a término–, o acaso una

simple exigencia de eludir la circularidad o la explicación *obscurum per obscurius* –aquella que, como señalaban Arnauld y Nicole (1662/1759: 297), prueba "una cosa incierta, por otra, que es tanto, ò mas incierta". Se trataría, en suma, de un reduccionismo mínimo: la doble asunción de que existen unas propiedades básicas, primitivas de cuanto existe y de que entre ellas no se cuenta la intencionalidad –o cualquier otra que sea la propiedad en cuestión, generalmente de índole psicológica, estética o ética. La idea de canon toma en Stich y Warfield, por ejemplo, la forma de una restricción sobre el rango de teorías aceptables:

> Perhaps the most common constraint placed on theories aimed at explaining the semantics of mental representations is the requirement that the theory be *naturalistic*. What exactly this "naturalism constraint" amounts to is a matter of considerable debate […]. But the basic idea is that semantic properties are not part of the basic building blocks of the universe and thus semantic facts should not be viewed as primitive or unanalyzable. Rather, the naturalism constraint insists, a theory of mental representation must provide some account of how semantic properties arise out of more basic non-semantic properties. (Stich y Warfield 1994b: 5)

En una discusión sobre la viabilidad del naturalismo ético, Harman (1977) aventura una definición más modesta –tanto que se hace irremisiblemente deudora de una aclaración del concepto de hechos naturales: "Naturalism as a general view is the sensible thesis that *all* facts are facts of nature" (Harman 1977: 17). En cuanto atañe a la naturalización de la intencionalidad, el criterio que de hecho parece operar es, en la práctica, bastante expeditivo: propiedades naturales serían aquellas que reconociese como elementales la física, o que pudieran construirse sólo a partir de éstas. Supone Fodor, por ejemplo, que:

> […] sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear on the list. But *aboutness* surely won't; intentionality simply doesn't go that deep. It is hard to see, in face of this consideration, how one can be a Realist about intentionality without also being, to some extent, a Reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really be something else. (Fodor 1987: 97)

Seguramente desde que George E. Moore encabezara con esa cita sus influyentísimos *Principia Ethica* (Moore 1903: 29) –uno de cuyos propósitos declarados era precisamente combatir en el ámbito de la filosofía moral lo que Moore dio en llamar la falacia naturalista: a saber, la confusión entre las propiedades que hacen que algo sea bueno y su bondad–, la controversia en torno al naturalismo ha quedado impregnada por el vago recuerdo del apologista y obispo anglicano Joseph Butler, y de su constatación retórica de que *todo es lo que es y no otra cosa* –una aseveración, "Everything is what it is, and not another thing", deslizada en el seno de una discusión sobre la relación entre la virtud y el desinterés (Butler 1726/2000: 323), que

el tiempo ha convertido en un extraño y desarraigado aforismo, extraviado en algún lugar entre el sentido común, la expresión coloquial del principio de identidad en lógica de predicados, y el sentimiento antifilosófico. Son las palabras de Butler, o de Moore, las que resuenan en el aforismo que cierra la reflexión de Fodor, así como en la escaramuza dialéctica entre Fodor y Searle que Stich y Warfield (1994b) citan como epítome de la controversia en torno al naturalismo. Cara a cara contesta Searle:

> You cannot reduce intentional content […] to something else, because if you could they would be something else, and they are not something else. The opposite of my view is stated very succinctly by Fodor: "If aboutness is real, it must really be something else." On the contrary, aboutness (i.e., intentionality) is real, and it is not something else. (Searle 1992: 51)[16]

Lo que aquí se confrontan, así pues, son intuiciones –aparentemente intuiciones brutas, aunque no cabe duda de que su genealogía intelectual ha de poderse rastrear– acerca de la pertenencia o no de la intencionalidad al inventario de propiedades elementales de la naturaleza. La cuestión es, en definitiva, si efectivamente queda trabajo por hacer –trabajo explicativo– una vez que se ha descrito la relación entre una actitud y una proposición (*cf.* Bechtel 1988: 73; Rodríguez 2001b: 225, *supra*), o sin llegados a ese punto sólo queda dar fe de que hemos hecho pie, de que hemos alcanzado lo que Wittgenstein (1969) veía como el lecho del río del lenguaje, o del conocimiento[17]. A ojos de Fodor, desde luego, ese instante en el que ya no procede seguir preguntando, no es en el que se encuentra quien postula la noción de actitudes proposicionales. Antes al contrario:

> Something has to be said about the place of the semantic and the intentional in the natural order; it won't do to have unexplicated "relations to propositions" at the foundations of the philosophy of mind.
>
> [...] It's not that there aren't propositions, and it's not that there aren't graspings of them; it's rather that graspings of propositions aren't plausible candidates for ultimate stuff. If they're real, they must really be something else. (Fodor 1985: 18)

---

[16] Sin embargo, el desarrollo de su argumentación parece llevar luego a Searle a enrocarse en una defensa de la irreductibilidad de la consciencia, dando en cambio por plenamente explicable en términos biológicos toda forma de intencionalidad que no conlleve consciencia. Sobre las dificultades que entraña para Searle mantener tal postura, y en general sobre la dificultad de deslindar el estudio de la intencionalidad del estudio de la consciencia, *cf*. Hermoso y Chacón (2000) y Chacón y Hermoso (2009).

[17] Un lecho del que formaban parte según él asertos como el que había hecho célebre G.E. Moore (1939, *infra*) –precisamente Moore– al asegurar que podía refutar tajantemente el escepticismo sobre la existencia de un mundo exterior a la mente levantando su mano y diciendo: "He aquí una mano" –con ánimo seguramente más templado que el de Diógenes de Sínope hacia algún seguidor de los eleatas cuando, según nos cuenta Diógenes Laercio (*Vidas* VI: 13), "[…] diciendo otro que no había movimiento, se levantó y se puso a pasear". Ya en las *Investigaciones*, al abordar la cuestión de qué significa seguir un regla, apunta Wittgenstein (1953: §217): "[…s]i he agotado los fundamentos, he llegado a roca dura y mi pala se retuerce" –desde el primer párrafo había dejado anunciado que "[…] las explicaciones tienen que terminar en algún lugar" (1953: §1).

Quedaría, pues, mucho trabajo por hacer: ni más ni menos que, como anunciaban Stich y Warfield (1994b: 5, *supra*), rendir cuenta de la manera en que tales propiedades semánticas pueden venir constituidas por otras que no lo sean, una tarea que, desde esta perspectiva, se revela además como de naturaleza fundacional, y a la que llamaríamos "naturalizar la intencionalidad".

> A traditional foundational problem in the theory of meaning is: *where do semantic properties come from*? The presupposition of this question is that the fact that a word (or a sentence, or whatever) means what it does can't be a brute fact. […] To put it in the standard philosophical jargon, semantic properties must *supervene on* non-semantic properties. There may be some properties that things *just* have; that they have for no reason at all. But if there are, they are the kinds of properties that basic physics talks about (like mass, charge, and charm). They certainly don't include the kinds of properties that semanticists talk about (like meaning *dog* or being a synonym of 'bachelor'). (Fodor y LePore 1991: 142-143)

Frente al tenue sarcasmo que se revela en los ejemplos elegidos por Fodor y LePore, el hecho de que el naturalismo se asume en ciertos ámbitos casi como un rasgo necesario de toda teoría bien formada queda acaso más claro a la vista de la mordacidad poco sutil de comentarios como el que abre el trabajo de Lycan titulado "Thoughts about Things":

> For years and years, philosophers took thoughts and *beliefs* to be modifications of incorporeal Cartesian egos. Happily, since early in the present century, it has become clearer that believers are complex organisms embedded in natural, physical environments and nothing (metaphysically) more than that; materialism in one form or another has prevailed ever since. (Lycan 1986b: 160)

Ahora bien, el propio Lycan concede –aunque sin abandonar el tono burlón– que la concepción naturalista de lo mental no está ni mucho menos madura:

> Yet the mere rejection of spookstuff has done little or nothing to illuminate the positive nature of thought and belief. (Lycan 1986b: 160)

Esa "iluminación" de la naturaleza del pensamiento bajo una perspectiva materialista debería permitirnos esquivar la concesión a la intencionalidad de idéntico carácter axiomático al que otorgamos a las propiedades físicas elementales. Pero dicha iluminación es, claro, lo que Brentano consideraba de todo punto impractible: de hecho, en la inviabilidad de esa labor cifraba Brentano la refutación del materialismo. La misma encrucijada –naturalización de la intencionalidad o renuncia al materialismo– queda cartografiada por Field en una de las primeras reflexiones filosóficas suscitadas por el papel del concepto de representación en el funcionalismo y la psicología cognitiva:

> So any materialist who takes believing and desiring at face value –any materialist who admits that belief and desire *are* relations between people and propositions– any such materialist must show that the relations in question are not irreducibly mental. Brentano felt that this could not be done, and since he saw no alternative to viewing belief and desire as relations to propositions, he concluded that materialism must be false. (Field 1978: 34)

Por aquilatar la exigencia planteada por Field, conviene apuntar que ésta no se vería cumplida con un gesto de reversión de una presunta hipóstasis previa de lo mental: no basta con negar la existencia de objetos mentales reiterando en cambio la de propiedades mentales, ni siquiera si a ello se añade la explicación de las razones que nos llevaron a la reificación ahora revocada; no basta, porque la idea de propiedades mentales es para el materialista –o eso mantiene Field– tan irremediablemente ininteligible como la de objetos mentales. Un dualismo psicofísico de propiedades, a la manera por ejemplo de Popper y Eccles (1984), no sería, así pues, contestación apta al reto de Brentano.

> Any interesting version of materialism requires not only that there be no irreducibly mental *objects*, but also that there are no irreducibly mental *properties*: the idea that, although people and certain higher animals do not contain any immaterial substance, nonetheless they have certain mental properties that are completely unexplainable in physical terms, is an idea that very few people who regard themselves as materialists would find satisfying. (Field 1978: 34)

Claro que todo esto se nos antoja bien distinto cuando el naturalismo se define como "[…] el desconocimiento del ser propio del conocimiento y su consiguiente 'naturalización' o 'cosificación', es decir, su asimilación a las cosas" (Llano 1999: 25). Entonces parece claro que la naturalización es –por decirlo con aire sartreano– una pasión inútil, un empeño infecundo que sólo puede apartarnos de la verdadera naturaleza de todo aquello que no es "las cosas". En efecto:

> Los trasuntos mentales –tanto intelectuales como imaginativos– no son susceptibles de recibir un tratamiento físico, ni siquiera convencionalmente psicológico. La óptica que permite acceder dificultosamente a ellos no está tamizada por las nociones de materia y causalidad, ni por tipo alguno de mecanicismo. En la mente nuestra se detecta algo *propio*, que no se halla en la realidad exterior y separada. La perspectiva adecuada para acercarse a ese tipo único de ser no es otra que la *intencionalidad*, la peculiaridad de una existencia que no consiste y persiste en sí misma, sino que remite a algo distinto de sí. (Llano 1999: 24)

Mirado desde esa misma animadversión a los principios del naturalismo, el proyecto toma un aspecto que, tomando al conductismo como ensayo paradigmático de llevarlo a cabo, puede bosquejarse así:

> La dificultad del conductismo, como de todo mecanicismo, reside, sin embargo, desde Hobbes a Russell, en cómo reducir la semántica a la sintaxis o, si se prefiere, una imagen o un afecto a movimientos implícitos, de estímulo y respuesta. (Domínguez 2003: 42)

Lo que se dirime, en suma, es si el desencantamiento del mundo –el *Enträthselung* en el que, según el relato de du Bois-Reymond (1898: 75, *infra*) se vería irremediablemente envuelto el científico investido de un conocimiento total del mundo físico– alcanza también los dominios del alma, o si éstos resultan inexpugnables incluso para el desdichado sabio fáustico.

Motivos para quebrar el hechizo

Vistas así las cosas, parecería que el sino de las propiedades psicológicas sólo puede ser quedar naturalizadas o ingresar en el inventario de propiedades elementales cuya existencia damos por axiomática. Bajo un prisma más abiertamente epistemológico, esto vendría en apariencia a ser tanto como decir: o bien todas aquellas teorías en las que tal o cual propiedad psicológica desempeñe un papel explicativo pueden ser satisfactoriamente reducidas a teorías en las que dicho papel quede cubierto por propiedades no psicológicas, o bien las teorías psicológicas en cuestión son autónomas y, al menos en tanto que irreductibles a ellas, se sitúan en pie de igualdad con las teorías físicas. Por supuesto, qué se entienda por reducir una teoría a otra, así como qué se entienda por propiedades básicas o elementales, no psicológicas, son asuntos espinosos que resurgirán en diversos momentos de este estudio. También lo es, desde luego, que la transición entre un discurso ontológico sobre propiedades elementales de la naturaleza y un discurso epistemológico sobre relaciones entre teorías –o, si se quiere, la propia distinción entre lo ontológico y lo epistemológico– sea tan tersa como acaba de esbozarse –sobre todo a la luz de que la convicción naturalista se acompaña a menudo, y rotundamente en Fodor, de un convencimiento igualmente firme respecto a que la psicología es de hecho una ciencia autónoma. No obstante, lo que reclama ahora una mirada más detenida es el aparente dilema entre reconocer rango primitivo, o elemental, a una determinada propiedad o quedar abocado a su naturalización. Se trata del mismo dilema que Horgan (1994) ve traslucir, implícito, en los ensayos de naturalización de la intencionalidad avanzados por Fodor o Cummins (*cf*. por ejemplo, Cummins 1983, 1989):

> So there is an underlying assumption behind philosophical projects such as Cummins's and Fodor's. It might be formulated this way:
>
> > Either (i) there are tractably specifiable non-intentional, non-semantic, sufficient conditions (or sufficient and necessary conditions) for something's being a mental representation with a specific content, or (ii) mental content is among the ultimate, fundamental, and unexplainable properties of things.
>
> This assumption is evidently so deeply ingrained that [it] is not clearly noticed *as* an assumption. [...]

> I think it is time for philosophers to notice it, to subject it to critical scrutiny, and to rethink the question of how to accommodate intentionality within a broadly naturalistic metaphysical worldview. (Horgan 1994: 309)

Ahora bien, la formulación de Horgan deja fuera una tercera posibilidad que está nítidamente presente en el pensamiento de Fodor –también sin duda en el de Cummins– con mucha mayor fuerza que la de una irreductibilidad de lo intencional que suele darse sencillamente por impensable: el eliminacionismo, el instrumentalismo, o alguna otras forma de antirrealismo acerca de la intencionalidad. El propio Horgan cita unas lúcidas palabras de la *Psicosemántica* de Fodor:

> […T]he deepest motivation for intentional irrealism derives […] from a certain ontological intuition: that there is no place for intentional categories in a physicalist view of the world; that the intentional can't be *naturalized*. (Fodor 1987: 97)

Dado que la imposibilidad de naturalizar la intencionalidad *sería* un motivo para considerarla irreal, la naturalización tiende a aparecer como un bálsamo para quien, como Fodor, rechace tal irrealidad. En su célebre "Semántica al estilo de Wisconsin", Fodor describía más detalladamente sus propios "motivos profundos" para anhelar una intencionalidad naturalizada:

> The worry about representation is above all that the semantic (and/or the intentional) will prove permanently recalcitrant to integration in the natural order; for example, that the semantic/intentional properties of things will fail to supervene upon their physical properties. What is required to relieve the worry is therefore, at a minimum, the framing of *naturalistic* conditions for representation. That is, what we want at a minimum is something of the form *"R represents S" is true iff C* where the vocabulary in which condition *C* is couched contains neither intentional nor semantic expressions. (Fodor 1984: 2)

Pero, obviamente, la imposibilidad de naturalizar la intencionalidad sólo es un motivo para considerarla irreal *una vez que* se ha repudiado la idea de que constituya una propiedad elemental, axiomática o, en las palabras de Horgan, "última, fundamental e inexplicable". De ahí que, como se ha dicho, la posibilidad de que la intencionalidad no exista tenga en la –por así decir– etiología del naturalismo, al menos en el caso de Fodor, una presencia mucho más vivaz que la posibilidad de que constituya un propiedad fundamental[18].

[18] Desde luego, si vernos llevados a negar la realidad de lo mental, comoquiera que entendamos esa negación, no se nos revela como una perspectiva inquietante, la tarea de dar cuenta de su imbricación entre las propiedades que aceptemos como naturales no parecerá, tampoco, perentoria. No es raro, así pues, que Stich (1992) ironice sobre la presunta urgencia de una teoría de la representación mental, cuya carencia supondría, a ojos de Dretske (1988: *x*) la renuncia a nuestra "[…] concepción de nosotros mismos como agentes" y, en consecuencia, para Fodor (1987: *xii*), "[…] la mayor catástrofe intelectual de la historia de nuestra especie" –o incluso, entre bromas y veras, "[…] el fin del mundo" (Fodor 1989: 77, Fodor 1990: 156, *infra*). La pregunta que Stich aborda es: "[…] ¿qué hace que el proyecto de

Lo que preocupa a Fodor, en suma, es que la carencia de una concepción naturalizada de la intencionalidad pueda esgrimirse, según hemos aprendido de Quine a hacerlo, como "[…] prueba de la falta de base de los giros intencionales y [de] la vaciedad de una ciencia de la intención" (Quine 1960: 280). Se trata, entonces de naturalizar la intencionalidad ni más ni menos que para salvarla de la quema: salvarla de calcinarse en una pira en la que Quine habría hecho ya arder la epistemología "tradicional" hasta que de ella quedara sólo psicología, y en la que el mismo fuego despojaría también a esa psicología –la "tradicional", la "ciencia de la intención"– de todos sus elementos volátiles, dejando sólo de ella el estudio de ciertas correlaciones entre estímulos y respuestas –cuanto en ella hay, diría Quine, de ciencia propiamente natural[19].

> Epistemology, or something like it, simply falls into place as a chapter of psychology and hence of natural science. It studies a natural phenomenon, *viz.*, a physical human subject. This human subject is accorded a certain experimentally controlled input –certain patterns of irradiation in assorted frequencies, for instance– and in the fullness of time the subject delivers as output a description of the three-dimensional external world and its history. The relation between the meager input and the torrential output is a relation that we are prompted to study for somewhat the same reasons that always prompted epistemology: namely, in order to see how evidence relates to theory, and in what ways one's theory of nature transcends any available evidence... But a conspicuous difference between old epistemology and the epistemological enterprise in this new psychological setting is that we can now make free use of empirical psychology. (Quine 1969: 82-83)

Frente a Quine, así pues, el propósito de Fodor es mostrarnos al menos cuál podría ser el esqueleto conceptual de una teoría naturalista de los fenómenos intencionales, de modo que pierda fuelle la tentación de prescindir de ellos en el desarrollo de la psicología científica –o, si en ésas estamos, de la epistemología[20]. Que la

---

generar tal teoría parezca tan urgente?"(Stich 1992: 347). Su concisa respuesta: "[…] there seems to be little doubt that many people think a theory of mental representation has a major role to play in the debate over eliminativism" (Stich 1992: 349), pese a que, a su juicio, renunciar a la naturalización de la intencionalidad, o incluso asumir su eliminación, acaso no dejara ninguna secuela en lo que atañe a su estatus ontológico (Stich 1992: 362, *infra*).

[19] En esta lectura de las dificultades del naturalismo se apoya Quine (1960: *ibid*.), además, para sostener que "[…] la tesis de Brentano de la irreducibilidad de los giros intencionales es inseparable de la tesis de la indeterminación de la traducción" (*cf. infra*).

[20] Con los años, como es sabido, el propio Quine (1990) acabaría por moderar su postura, tratando de incorporar a su proyecto de naturalización alguna suerte de reconstrucción del elemento normativo de la epistemología tradicional. Que tal reconstrucción sea factible ha sido, desde luego, objeto de intenso debate. Pero también entre algunos de los más severos críticos de Quine es fácil encontrar el rastro de un consenso naturalista que, puesto que alcanza incluso al ámbito de la epistemología, será sin duda más robusto aun en el de la psicología. Así, por ejemplo, Kim (1988: 399) escribe: "[…] if a belief is justified, that must be *because* it has certain factual, non-epistemic properties […]. That it is a justified belief cannot be a brute fundamental fact […]. [It] must be grounded in the factual descriptive properties of that particular belief". El primado de Kim en la reciente defensa de un estricto fisicalismo frente a lo que él contemplaría como veleidades funcionalistas, que se abordará en detalle *infra*, es

intencionalidad no es una propiedad fundamental de la naturaleza, y que no puede por tanto quedar inexplicada en una teoría que se pretenda naturalista, sencillamente se da por sentado. Buena parte de la trayectoria filosófica de Fodor puede entenderse entonces, precisamente, como un esfuerzo por hacer explícito el modo en que la psicología pudiera preservar su autonomía en el seno de la ciencia sin renunciar a las aspiraciones naturalistas o, dicho de otro modo, por forjar un naturalismo sin reduccionismo –tan es así que el proyecto de Fodor, cuando no su particular manera de llevarlo a cabo, ha acabado casi por confundirse con la tarea misma de dotar de fundamentación filosófica a la aproximación cognitivista a la psicología. Así, por citar sólo un par de ejemplos, "Materialismo sin reduccionismo" es de hecho el título de un escrito de Boyd (1980, *infra*) en el que, con parecido propósito, se revisa la noción de clase natural manejada por el fisicalismo; idéntico espíritu anima los primeros capítulos de Pylyshyn (1984, *infra*), donde la labor de desgajar naturalismo y reduccionismo se hace descansar sobre el papel crucial que el tipo de vocabulario empleado para describir un fenómeno desempeña a la hora de delimitar el repertorio de generalizaciones que podemos capturar acerca de él.

Particularmente sensible a los motivos de Fodor se ha mostrado Moya (1994) al apuntar la conveniencia de reconsiderar radicalmente nuestra idea de las relaciones entre naturalismo y humanismo:

> En general, se ha tendido a dar por supuesto que los intereses del humanismo eran mejor servidos por la insistencia en la autonomía e irreductibilidad del mundo del espíritu [...]. Sin embargo, este supuesto [...] no puede ser considerado como evidente. De hecho, puede que sea falso. Humanismo y naturalismo podrían no ser perspectivas antitéticas. (Moya 1994: 228)

Es más:

> En este contexto, el proyecto de naturalización de la intencionalidad, la comprensión de la intencionalidad en términos de propiedades originariamente no intencionales, aceptables para una concepción materialista de la realidad, aparece, con cierto aire de paradoja, como un aliado del humanismo [...]. (Moya 1994: 229-230)

No en vano, también Moya hace descansar esta idea sobre el giro impreso por Quine a las tesis de Brentano, y también, como Fodor, encuentra evidente a la razón que la intencionalidad no puede ser una propiedad fundamental:

> [...A]sí pues, *debe* ser posible analizar la intencionalidad en sus componentes no intencionales, reconstruir conceptualmente la conjunción en interacción de factores que ha dado lugar a la intencionalidad en el mundo natural. Y si esto no es posible, o bien hemos de aceptar, contra toda razón, que las propiedades intencionales son inexplicables y últimas, o bien hemos de reconocer que son ficciones carentes de realidad. (Moya 1994: 233)

---

desde luego plenamente congruente con este su naturalismo epistemológico, que, por lo demás, no es óbice para su discrepancia con el proyecto original de Quine.

Dista de estar claro, desde luego, que ese análisis que "*debe* ser posible" lo sea de hecho, y lo cierto es que en esa labor que ha quedado antes descrita como un intento de iluminación de la naturaleza del pensamiento bajo una perspectiva materialista, el historial de ensayos es a día de hoy más bien desalentador. Una y otra vez, bajo la apariencia de un *lógos* escuetamente natural –de acuerdo, eso sí, con cambiantes criterios de austeridad, o de naturalidad– ha ido velándose la apelación decisiva a hechos intencionales que desbarataba el trazado entero de la teoría. Como sumarísimamente registra Moya:

> Aunque la expresión "naturalización de la intencionalidad" es bastante reciente, el proyecto designado por ella cuenta con una larga historia, una historia sembrada de fracaso. El atomismo griego, la teoría de las pasiones de Spinoza, la antropología de Hobbes, el materialismo de Condillac, son sólo algunas muestras de este repetido fracaso. En nuestro siglo, el conductismo y el materialismo de la identidad de propiedades son programas de reducción naturalista de la intencionalidad que han sido prácticamente abandonados. La plausibilidad aparente que han podido tener algunos de estos proyectos se debe, en gran parte, al hecho de que las categorías intencionales han sido implícitamente presupuestas bajo una superficie no intencional. (Moya 1994: 233)

La cuestión se formula en Horgan (1994) –recordemos– a modo de dilema. Con la vista puesta en la alternativa entre *(i)* que la intencionalidad sea susceptible de quedar explicada en términos asumibles para el naturalista –es decir, que puedan especificarse, sin recurso a vocabulario intencional o semántico, un conjunto tratable de condiciones al menos suficientes[21] para que algo constituya una representación mental con un contenido específico– y *(ii)* que sea una propiedad fundamental, cuya postulación revista un carácter axiomático, la escapatoria bosquejada por Horgan pasa por convenir en tres premisas de diversa litigiosidad. En primer lugar, habría que dar la razón a Horgan en cuanto a que nuestra reticencia a adjudicar a la intencionalidad un estatus metafísico básico tenga que ver, al menos en parte, con la convicción de que las propiedades intencionales *supervienen* en propiedades físicas –es decir: de que es forzoso que exista alguna variación en las propiedades físicas de un organismo para que exista alguna variación en sus propiedades intencionales. En segundo lugar, sería preciso conceder que tal superveniencia no implica de suyo que existan condiciones como las especificadas en *(i)*, puesto que "[...a]lthough a physical supervenience base might always exist for any manifestation of aboutness, in general any adequate non-intentional, non-semantic characterization of the supervenience base might be enormously baroque and complex" (Horgan 1994: 309), y, por tanto, contra *(i)*, escasamente *tratable*. Por último, deberíamos seguir a Horgan en su

[21] Con propiedades *al menos* suficientes parecería darse por satisfecho Horgan, aunque, como veremos, Stich (1992) cifra la viabilidad de proyecto de naturalización de la intencionalidad en la formulación de condiciones suficientes y necesarias. También Fodor (1984: 2, *supra*) especifica con rotundidad que "what we want at a minimum is something of the form *'R represents S' is true iff C* where the vocabulary in which condition *C* is couched contains neither intentional nor semantic expressions".

convicción de que si la relación de superveniencia encontrada entre las propiedades intencionales y las propiedades físicas resulta ser, en efecto, de tal suerte que no puedan proporcionarse condiciones como las descritas en *(i)*, ello no sería suficiente para dar por cumplida la naturalización de la intencionalidad, ya que la propia refractariedad a la explicación de la relación de superveniencia –la susodicha "intratabilidad" del conjunto de condiciones que la describen– constituiría de por sí un poderoso motivo para considerar la intencionalidad una propiedad fundamental:

> [...f]or, if certain interlevel supervenience facts are themselves *sui generis* and unexplainable, then the supervening properties will thereby qualify for inclusion on the list of ultimate and irreducible properties of things –supervenience notwithstanding. (Horgan 1994: 310).

Además, Horgan parece convencido de que aceptar su propuesta exigiría concordar, a la manera de una cuarta premisa, en cierta interpretación de la investigación psicológica sobre categorización, que ha arrojado como fruto nociones tales como las de "prototipo" y "ejemplar", y de acuerdo con la cual hay de hecho razones para el escepticismo en cuanto a la existencia de condiciones que, bajo los criterios fijados en *(i)*, puedan dar cuenta de la existencia de una representación mental con un contenido determinado, razones que tendrían que ver en realidad con la imposibilidad de definir mediante condiciones suficientes –menos aún necesarias– los conceptos humanos en general[22].

Nos hallaríamos con todo esto, piensa Horgan, ante una variante de la idea de propiedad fundamental que no vulneraría las convicciones naturalistas, al menos en tanto que reconoce la superveniencia de la propiedad a la que se da tal carácter fundamental en ciertas propiedades físicas[23]. El carácter primitivo, elemental que

[22] La evidencia al respecto provendría de la psicología del pensamiento: sus fuentes principales serían los trabajos de Rosch (1973, 1975, 1978), primero en torno a la idea de *prototipo* y más adelante en torno a la de *ejemplares* cuya similitud con un ítem dado, evaluada de forma tácita, decide la pertenencia del ítem a una categoría, y particularmente los trabajos de Barsalou (1987) y Rips (1989), según los cuales "[…] los sujetos construyen conceptos 'al vuelo', en respuesta a la situación en la que surge la necesidad de categorizar" (Stich 1992: 353). La relevancia de la investigación sobre categorización para la polémica entre realismo y eliminacionismo respecto a lo intencional, o lo mental en general, ha sido puesta también de manifiesto, entre otros, por Stich (1994: 352-353, *infra*), en un razonamiento del que parece heredero el hilvanado por Horgan.

[23] Aunque, como bien ha recalcado Feldman (2001), "[…i]t [is] difficult to determine whether supervening on what is natural is sufficient for naturalness". A primera vista, no resulta descabellado que pensar que el carácter natural de una propiedad y su superveniencia en propiedades físicas puedan ir ligados: las propiedades en cuestión "[…] do not float free, they are not autonomous, they are not brute facts, they are anchored in the physical world. That seems like a good reason to conclude that they are natural facts" (Feldman 2001: §4). Sin embargo, cuando la cuestión del naturalismo se traslada del terreno psicológico, o epistemológico, al terreno ético, vincular superveniencia y naturalización arroja un resultado inesperado: el consenso naturalista parece extenderse también a la ética, incluso al trabajo de declarados antinaturalistas como G.E. Moore. Si esto, como Feldman sugiere, es un motivo para dudar del análisis de la noción de propiedad natural en términos de superveniencia, queda al menos el alivio de que es un motivo débil. Si, como advierte Kim, es un

reviste la propiedad en cuestión no quedaría cifrado en el quebrantamiento de ningún principio de superveniencia sino, antes bien, en la imposibilidad de expresar las propiedades físicas relevantes como un conjunto tratable de condiciones necesarias o, *a fortiori*, necesarias y suficientes. En hacer inasequible tal resultado confluirían, por un lado –diríase: el lado epistemológico–, el hecho de que las nociones empleadas en el intento –la de intencionalidad, las de las propiedades físicas en cuestión– son al fin y al cabo el fruto de procesos de categorización y formación de conceptos tan poco proclives a redundar en definiciones *more aristotelico* como cualquier otro proceso cognitivo humano –salvo, naturalmente, la propia labor, explícita y *ad hoc*, de definición por género y diferencia–, y, por otro lado –diríase: el ontológico–, el presumible carácter último e inexplicable, *sui generis*, de algunos aspectos de la propia relación de superveniencia. Cuál sea la relación entre ambas fuentes de intratabilidad es un asunto que Horgan no aborda.

Sea como sea, la conclusión que Horgan deriva de estas premisas es que podrían ser simultáneamente falsos ambos cuernos del dilema de partida, *(i)* y *(ii)* –esto es, que por mucho que eventualmente resultara inviable detallar un conjunto tratable de condiciones suficientes para que un fenómeno psicológico tenga tal o cual contenido intencional, o que tal empresa se revelara inviable por razones de principio, bien podría la intencionalidad, con todo, desmerecer ese carácter de propiedad última, fundamental e inexplicable, o, más bien, merecerlo únicamente en un sentido en que el conflicto con el naturalismo se desvanece. Si Horgan está en lo cierto, claro, la disyuntiva descrita como presupuesto del pensamiento de Cummins y Fodor no es tal disyuntiva –no hay disyuntiva donde ambos términos pueden ser falsos, pues la falsedad de ambos implicaría la de la disyunción. Por supuesto, *(i)* y *(ii)* serían ambas falsas en caso de que fuese verdadero el tercer término elidido en la disyuntiva que Horgan plantea: el eliminacionismo. Lo que Horgan pretende, en cambio, es que, siendo ambas sean falsas, el naturalismo quede reivindicado, o, dicho de otro modo, que la falsedad de ambas sea aceptable desde criterios naturalistas y no conduzca, por tanto, al eliminacionismo.

Pero el argumento es débil. Dado que la primera premisa no hace sino esbozar una descripción expresamente parcial de los motivos por los que *tendemos a entender* la intencionalidad como ajena al ámbito de las propiedades fundamentales, parece razonable darla provisionalmente por buena. La clave para asegurar la segunda residiría, por lo que parece, en la incierta noción de que las condiciones especificadas resulten "tratables", por oposición a "barrocas". A primera vista, es difícil negar que podría darse una superveniencia de las propiedades intencionales en las propiedades físicas que no pudiera expresarse de forma "tratable" en términos de condiciones físicas suficientes para que se den determinadas propiedades intencionales, pero es una obviedad que la cuestión depende de qué se entienda por "tratable".

---

signo de la equivocidad de la noción de naturalismo, que se emplearía en un sentido en epistemología y en otro más exigente en ética, parece que Horgan puede despachar las correcciones terminológicas o conceptuales a que haya lugar sin que su argumento languidezca.

La interpretación más natural es que "tratable" es una propiedad relacional que implica al conjunto de condiciones suficientes y a nuestras capacidades cognitivas: *¿podríamos nosotros* entender tal conjunto, aprehenderlo, operar con él? Fuera cual fuera la respuesta, quedaría pendiente aclarar si las capacidades cognitivas que se han puesto en juego vienen o no mediadas por nuestros conocimientos –en los términos propios del estudio psicológico sobre inteligencia, si se trata de capacidades cristalizadas o fluidas–, es decir, si el conjunto se nos hace intratable porque carecemos de los conocimientos precisos o porque nos faltan las competencias básicas precisas. En el primer caso, el argumento de Horgan conduciría como mucho a una constatación de la indigencia de nuestro saber, perfectamente sensata pero sin más consecuencias que la de confirmar que la naturalización de la intencionalidad es, como ya sabíamos, un proyecto inacabado. En el segundo caso, entendida la intratabilidad del conjunto de condiciones físicas necesarias para que se de un cierto fenómeno intencional como una limitación de principio al venidero avance de nuestras indagaciones, son varios los escollos con los que topamos. No el menor de ellos, sin duda, es que Horgan no ha proporcionado argumento alguno para convencernos de que existen tales limitaciones. Si bien eso hace razonable pensar que no es ésta la interpretación de la idea de intratabilidad que Horgan tiene en mente, conviene tener presente que otros, en la estela de Huxley (1866) y Tyndall (1871) sí han ensayado el camino que Horgan elude y que –recordemos– se destila en el sombrío augurio de Du Bois-Reymond (1872, *supra*): "*Ignorabimus!*". Así, la reiteración de las tesis de Huxley o Tyndall es patente –si bien, como ha señalado Tennant (2007: 751), no es fácil dilucidar si existe un hilo ininterrumpido de influencia intelectual o una mera reaparición de ideas ya ensayadas tiempo atrás– en planteamientos como el que ha venido enarbolando McGinn (1989, 1999), según el cual:

> […] we are cut off by our very cognitive constitution from achieving a conception of that natural property of the brain (or of consciousness) that accounts for the psychophysical link […]. We should […] be alert to the possibility that a problem that strikes us as deeply intractable, as utterly baffling, may arise from an area of cognitive closure in our ways of representing the world. (McGinn 1989: 350-352)

Las palabras de Huxley, en 1872, expresaban cuando menos un ánimo muy semejante en cuanto a las expectativas de deshacer el nudo que enlaza a la mente y el cuerpo:

> […] the reason of the connection between the molecular disturbance and the psychical phenomenon appears to be out of the reach, not only of our means of investigation, but even of our powers of conception. (Huxley 1866/1872: 301)

Mayor esmero pondría du Bois-Reymond en sostener la desalentada conclusión que Huxley daba por evidente, fabricando a tal fin un evocador experimento imaginario en el que un ser dotado de "conocimiento astronómico" (du Bois-Reymond 1898: 75)

–es decir, según su definición, que abarca la totalidad de las observaciones correctas que resulten precisas y está exento de toda dificultad teórica– se ve incapaz de entender el surgimiento de la vida consciente, incluso en lo que de otro modo sería su omnisciencia. Con todo, el propio du Bois-Reymond (1898: 69) admitiría que dicha conclusión siempre le pareció "una verdad trivial". En la joven estirpe del reconocimiento de esa supuesta trivialidad trataría el ya anciano fisiólogo de concitar no sólo a Tyndall, sino a Julian Offray de la Mettrie e incluso a Leibniz[24]. Sin embargo, el trasfondo fáustico de la preocupación por los límites del conocimiento, que cristaliza en la idea de "conocimiento astronómico", haría cobrar al *ignorabimus* una fuerza que difícilmente hubiera podido alcanzar un siglo atrás, antes de que Goethe (1808, 1832) revigorizara el viejo mito del sabio y Mefistófeles. Pero no mucho después la intuición que para du Bois-Reymond daba cuerpo a una obviedad había quedado ya marchita a ojos de Edward Lee Thorndike, quien no dudaba de que:

> […i]f we had perfect knowledge of the entire history of a man's brain, if we could from second to second see just what is going on in it, we should find in its actions and consequent changes the parallel of his life of thought and action. (Thorndike 1905/1912: 168)

En esa confianza en los frutos venideros de la investigación, igual que en tantas otras cosas, el pensamiento de Thorndike parece marcar la senda que recorrería la psicología científica en las décadas posteriores, poco afín a la desesperanza. Al contrario: ya un año antes de la publicación del manifiesto conductista de John B. Watson (1913), la tesis de que la conciencia, las voliciones o los sentimientos –el residuo incognoscible que vislumbrara du Bois-Reymond– no son más que entidades mitológicas, a las que se comparaba con fantasmas, dioses o demonios, quedó

[24] En las páginas de *El hombre máquina* cree du Bois-Reymond (1912: 528) haber encontrado la tesis de que "[…] nunca comprenderemos cómo puede pensar la materia". Quizá se refiriese a la afirmación de que "[…] los lazos que reinan entre la causa [cerebral] y los efectos [sobre el comportamiento de los animales] [… son] una especie de armonía que los filósofos no comprenderán nunca" (de la Mettrie 1748: 18). No parece haber en *El hombre máquina*, sin embargo, una expresión más rotunda de un agnosticismo como el alentado por Huxley, Tyndall o du Bois-Reymond –ésta no lo es tanto, dado el aire despectivo con que de la Mettrie, médico de profesión, acostumbra a referirse a *los filósofos*–, ni rastro más preciso de la idea que du Bois-Reymond cita informalmente. De la conocida parábola leibniziana del molino pensante (Leibniz 1715: §17) extrae asimismo du Bois-Reymond la idea de que no cabe una explicación mecánica de la percepción, puesto que nada que pudiéramos contemplar al pasear por el interior de una máquina dotada de pensamiento y sentimiento, pero del tamaño de un molino, nos ayudaría a entenderla –si bien lo hace dejando de lado la resolución a la que ello condujo a Leibniz, y que se halla en el corazón de la *Monadología*: que la explicación de la percepción "[…] hay que buscarla en la sustancia simple", siendo la percepción, precisamente lo único en que "[…] pueden consistir todas las *acciones internas* de las sustancias simples".

Parece claro, así pues, que las intuiciones de Leibniz descansaban más bien del lado de la idea de que lo mental sea una propiedad última de la realidad. Pese al halo de radicalidad materialista que rodea a su nombre desde que los sucesivos escándalos provocados por sus publicaciones lo forzaran a exiliarse en Holanda y luego en Prusia, lo mismo puede decirse sobre de la Mettrie: *cf.* de la Mettrie (1748: 246, *infra*).

bosquejada por Max F. Meyer (1912: 367, *cf*. también Meyer 1908: 360-361) –quien dicho sea de paso, en 1911 había dado a la imprenta *The Fundamental Laws of Human Behavior*, un rotundo alegato en favor de la reducción de toda la psicología a explicaciones neurológicas[25] que incorporaba la idea, heredada de Lazarus Geiger y que luego sería medular al conductismo, de que el pensamiento no es más que *innere Sprache*. Trazar semejanzas entre los conceptos psicológicos y la mitología ha venido siendo, al menos desde entonces, un divertimento predilecto del eliminacionismo.

De cualquier forma, la conclusión de du Bois-Reymond, así vista, no parece diferir mucho de una renuncia parcial a la propia idea de conocimiento que –máxime a falta de *argumentos*, como se presentaría en Horgan– no sería descabellado, como ya hacía Cajal (1897, *supra*), considerar *todavía* prematura[26]. Aun si recabasemos argumentos a todas luces convincentes respecto a nuestra propia incapacidad de entender la naturaleza de la intencionalidad, es difícil, desde luego, intuir entre los rasgos de una renuncia de ese calado alguno que la hiciera grata al naturalismo, cuya reivindicación Horgan pretende. Pero, por otra parte, si la noción de intratabilidad no ha de entenderse, tal como invita a hacerlo, en un sentido relacional –o no al menos en un sentido que ataña a nuestras capacidades cognitivas–, entonces es patente que la labor de especificar en qué sentido debe entenderse está por hacer, y también que la fuerza del argumento de Horgan pende de ese hilo.

---

[25] Con algo más de cuidado que el que luego podrá advertirse en Skinner, Meyer diferencia aquellos conceptos subjetivos para cuyas referencias podrán hallarse, con el avance de la investigación, correlatos nerviosos específicos, y que por tanto quedarán sujetos a un programa de traducción al vocabulario objetivo de la neurología, de aquellos otros conceptos subjetivos para los que "[…] no nervous correlate can be found", los cuales, Meyer dice atreverse a presagiar, "[…] are the very ones which are superfluous [… and] can be spared from our descriptions of mental life in man and animals" (Meyer 1912: 371).

No cabe por tanto imputar a Meyer oscilación alguna entre posiciones reduccionistas y eliminacionistas, en tanto que deja articulado un criterio firme que sirve de fiel de la balanza; tampoco parece acertado atribuirle un eliminacionismo sin matices, como parece hacer Cordeschi (2002: 259), ni considerarlo de todo ajeno a la tentación eliminacionista, como plantea Wozniak (1993) tras un estudio algo más pausado de su obra cardinal, *The Fundamental Laws of Human Behavior*. Es cierto, desde luego, que Meyer propugna la traducción del vocabulario psicológico al de la neurología, pero no lo es menos que se muestra taxativo respecto a que "[…] those terms of mental function, for which no nervous correlate can be found […], can be spared from our descriptions of mental life in man and animals" (1911: 371). La figura de Meyer, en relación con la de su discípulo Albert P. Weiss, aparece de nuevo *infra*; en Weiss, la sombra de la eliminación es, como veremos, también patente.

[26] Como prematuro resulta a ojos de Dennett (1990: 37) no ya abandonar el anhelo de desplegar un día una explicación enteramente naturalizada de lo mental, por haber cedido en alguna de sus formas al espectro del *ignorabimus*, sino incluso cualquier tímido coqueteo con el dualismo, puesto que "[…] given the way that dualism wallows in mystery, accepting dualism is giving up". Es, pues, el convencimiento –sin duda compartido por un buen número de científicos y filósofos embarcados en el proyecto de naturalización– de que adoptar una posición dualista es tanto como alojar un milagro en nuestras explicaciones, lo que lleva a Dennett (1990: *ibid.*) a ceñirse a "[…] the apparently dogmatic rule that dualism is to be avoided *at all costs*", aun a falta de un argumento decisivo en su contra.

Pensar sin pensar

El germen del eliminacionismo es la idea de que nuestro discurso cotidiano sobre lo mental es de naturaleza teórica. Ya Carnap (1956: 70), cuando trataba de reformular la articulación entre el vocabulario teórico y el vocabulario observacional que había formado el núcleo del positivismo lógico –y que él mismo daba entonces por insostenible–, había englobado en el ámbito del vocabulario teórico los conceptos psicológicos coloquiales, además de los acuñados en el marco de la investigación psicológica. Eso parecía a todas luces un cierto alivio de los exigentes requisitos que la ortodoxia positivista y operacionalista había fijado para el empleo de conceptos psicológicos, y que el conductismo había hecho suyos: Wilfrid Sellars (1956: 316), por ejemplo, anunciaba que el conductismo se había mostrado "indebidamente restrictivo", pues nada impide considerar que las experiencias internas no son hechos "[…] 'empíricos' en el simple sentido de que son *teóricos*" (Sellars 1956: 318). Pero la aparente distensión de la actitud ante los conceptos psicológicos acabaría por propiciar un ademán si cabe más radical que el de exigir su estricta definición operacional: el de abogar por su total impugnación, por su indefectible abandono.

El mismísimo Smart (1967), consagrado desde la publicación de "Sensations and Brain Processes" (1959) como adalid del proyecto de identificar nuestros estados mentales con estados del sistema nervioso central, se vería tentado de asumir que lo mejor que podía hacerse con el vocabulario en que se articulan los enunciados con que nos atribuimos sensaciones era prescindir de él. La vacilación de Smart se produce en el contexto de su preocupación por las críticas que, sobre todo a manos de Bradley (1964), recibiera su propuesta de análisis temáticamente neutral de dichos informes de sensaciones (*cf. infra*), que era tildada de circular y, por tanto, de irremisiblemente mentalista. Con toda seguridad, no obstante, hicieron mella también en sus convicciones las criticas de Stevenson (1960), quien esgrimía el principio de indiscernibilidad de los idénticos para argumentar que si una sensación y un proceso cerebral dado son, como quiere Smart, estrictamente idénticos, entonces el proceso cerebral ha de poseer cuantas propiedades pudiéramos con verdad atribuir a la sensación –y no sólo viceversa. El propio Stevenson apunta que si Smart quiere evitar que la tesis de identidad psicofísica lo aboque así a un dualismo de propiedades, la única escapatoria que en la práctica le queda es el eliminacionismo: negar la existencia de las propiedades psicológicas. Pero, por supuesto, negar que existan propiedades psicológicas viene a ser tanto como negar que existan sensaciones –ya que, en el esquema de Smart, el único modo que tenemos de identificar las sensaciones es por medio de sus propiedades psicológicas–, y negar que existan sensaciones es tanto como negar que existan procesos cerebrales idénticos a tales sensaciones, es decir, negar la tesis de partida. No tardó Smart (1961: 93) en dar la razón a Stevenson: la tesis de identidad psicofísica implica la inexistencia de propiedades psicológicas, pues las supuestas propiedades psicológicas se revelan en realidad como propiedades físicas. O, dicho de otro modo,

"Paul Feyerabend may be right in his contention that common sense is inevitably dualistic, and that common sense introspective reports are couched in a framework of a dualistic conceptual scheme" (Smart 1967: 91)

Si bien Smart se repondría pronto de los cantos de sirena de la eliminación –o al menos creería hacerlo: en Smart (1972) presenta ya sus veleidades eliminacionistas como un error pasajero–, otros perseverarían en explorar esos derroteros, ya hollados por Quine (1960): es el caso no sólo de Feyerabend (1963), sino también de Rorty (1965) –quien distinguiría en el seno del fisicalismo entre una *forma de traducción*, según la cual la relación entre el vocabulario psicológico y el neurológico se asemejaría a la que se da entre "flujo calórico" y "energía cinética media", y una *forma de eliminación*, más coherente a su entender aunque poco recomendable en sentido pragmático, según la cual la relación es más bien como la existente entre "posesión demoníaca" y "crisis epiléptica"– e incluso de Feigl (1967: 141), que se muestra dispuesto a conceder a Smart –al Smart de 1967– y a Feyerabend que "[…] within the conceptual frame of theoretical natural science genuinely phenomenal (raw feel) terms have no place". En pocas palabras, so pena de sobresimplificación: el flujo calórico existe, *es* la energía cinética media; las posesiones demoníacas no existen, *no son* crisis epilépticas. Los conceptos teóricos provenientes de teorías erróneas, como el de posesión demoníaca, no son susceptibles de traducción a conceptos equivalentes en el marco de teorías verdaderas; en consecuencia, lo que procede es su total repudio: su eliminación. Para Feyerabend, Rorty o Feigl, al igual que para Smart (1967), el caso de las sensaciones crudas evocaba más al de las posesiones que al del flujo calórico.

Naturalizar la intencionalidad –bien lo sabemos– sería salvarla de las fauces de la eliminación, que ya devorara las posesiones demoníacas, el flogisto o el magnetismo animal. La clave –claro– es en qué consista exactamente naturalizar la intencionalidad. Seguramente evocando el catálogo de propiedades últimas e irreductibles que los físicos imaginados por Fodor (1987: 97, *supra*) se dedicarían a compendiar –y que el propio Stich citaría, como hemos visto, en Stich y Warfield (1994b)–, y anticipando el papel crucial que concede Horgan (1994: 309, *supra*) a la formulación de condiciones necesarias y suficientes como requisito para dar una propiedad por cumplidamente naturalizada, Stich parte de un canon de naturalización cuyo carácter abiertamente difuso pretende ser un reflejo del que predomina en el debate al respecto:

> [...W]hat does it take for an account of mental representation to be *naturalistic*? Though I know of no one who has offered a detailed answer to this question, the literature strongly suggests that those who want a naturalistic account of mental representation want something like a definition –a set of necessary and sufficient conditions– couched in terms that are unproblematically acceptable in the physical and biological sciences. (Stich 1992: 362)

Como haría Horgan (1994), Stich da prácticamente por hecho el fracaso de ese proyecto, al menos así entendido. Cosa muy distinta es que dicho fracaso haya de

provocar las terribles secuelas que Fodor o Dretske pregonan. Antes al contrario –piensa Stich:

> Whether an appropriately naturalistic account of mental representation can be given is, of course, very much an open question. My own guess, for what little it's worth, is that the project is quite hopeless. However, in contrast with Fodor and many others, I am inclined to think that very little hangs on the matter. (Stich 1992: 362)

La razón es que de la imposibilidad de culminar el proyecto de naturalización de la intencionalidad –al menos, de nuevo, así entendido– no se sigue a juicio de Stich ninguna conclusión relevante respecto a su realidad o irrealidad; tampoco, parece leerse entre líneas, respecto a otras facetas de su estatus metafísico. A modo de colofón de su argumento aduce Stich un par de ejemplos, tomados de disciplinas tan acreditadas como la lingüística y la etología, en los que naturalización y "respetabilidad empirica" no parecen ir de la mano:

> […] To see why, we need only consider a few examples. Let's begin with the notion of a phoneme. What is it to be a /p/ or a /b/? […D]espite many years of sophisticated research, there is currently no naturalistic answer available. […] Much the same point could be made about lots of other notions of unquestionable scientific utility. There is no naturalistic account of grooming behavior in primate ethology. Nor is there a naturalistic account of attack behavior in stickleback ethology. But surely it would simply be perverse to deny the existence of grooming behavior, simply because we can't define it in the language of physics and biology. Suitably trained observers can detect grooming behavior (or phonemes) with impressively high intersubjective reliability. And that, I would urge, is more than enough to make those notions empirically respectable. To demand more –in particular to demand that the notions in question can be "naturalized"– seems unmotivated and silly. (Stich 1992: 362)

Existen conductas de acicalamiento entre los primates, y existen fonemas. Incluso realidades más endebles, cuya existencia parece mucho más proclive a deshilachársenos entre los dedos, se diría que merecen una suerte algo menos inclemente que la que corrió el magnetismo animal. El ejemplo de las constelaciones sirve a Braddon-Mitchell y Jackson (1996) para distinguir en el seno del eliminacionismo una versión si se quiere más radical, que en la estela de Meyer (1908: 360-361, 1912: 367; *supra*) asemeja creencias y deseos a flogisto o posesión demoníaca, y otra más moderada, que considera a los conceptos de la psicología ordinaria afines, salvando las distancias, a otros como *Orión* u *Osa Mayor*:

> It was once thought that these constellations had some special explanatory significance. We now know that although they certainly exist, they have no significance for astronomy or science in general. [...] Beliefs and desires might be like constellations. (Braddon-Mitchell y Jackson 1996: 253)

La cuestión de fondo, como atinadamente apuntan Braddon-Mitchell y Jackson, es si a la hora de decidir si algo pasa a integrar o no un hipotético inventario científico de

todo cuanto existe conviene exigirle la condición de constituir una clase natural –y, en consecuencia, qué se entienda por clase natural:

> *If* by a natural kind is meant simply a kind that plays an important role in a predictive and explanatory science, then many functional kinds –vitamins, thunderstorms and terminal illnesses– will count as natural kinds. And so will beliefs and desires, because the functional roles they play are important for predicting and explaining behavior. But constellations won't count. Of course, if, as is usual, "natural kind" is taken to mean something more demanding: intrinsic similarity, common aetiology or whatever, beliefs and desires may not count. (Braddon-Mitchell y Jackson 1996: 253-254)

Pero aclarar qué entendamos por clase natural sería ocioso, en este contexto al menos, si como sostienen Braddon-Mitchell y Jackson resultara errónea la tesis de que sólo los conceptos que definen clases naturales están legitimados para formar parte del discurso científico.

En todo caso, la legitimidad de su imbricación en el aparato de la ciencia, e incluso de su propio estatus en el seno de lo existente, ha quedado en entredicho incluso para conceptos en apariencia mucho más incontrovertibles que los que Braddon-Mitchell y Jackson aducen como *argumento ab exempli* –vitaminas, tormentas eléctricas, enfermedades terminales. La noción de *enfermedad* es un caso destacado: ya en fecha tan temprana como para una historia del eliminacionismo, o de sus raíces positivistas, es 1923, tomaban nota Ogden y Richards (1923: 67), en el curso de una bella reflexión acerca del poder de las palabras y de su capacidad para hacernos presos de fantasías metafísicas –*vocem proferre et nihil concipere*–, de las osadas advertencias que el epidemiólogo Francis G. Crookshank lanzaba en uno de los suplementos que acompañaban al tratado: "[…] bajo la influencia de ciertas escuelas de pensamiento, y ciertos hábitos de expresión, nos hemos acostumbrado a hablar y escribir como si una enfermedad fuera un objeto natural", existente "[…] *in rebus Naturae* […]", (Crookshank 1923: 358), pero "[…] no es probable que se realice ningún progreso importante en el dominio de la Medicina, en tanto no se abandone la creencia en la existencia real de las enfermedades", que el propio Crookshank atribuye, entre otros motivos, a la "tiranía de los nombres" (Crookshank 1923: 359) –expresión de nítidos ecos fregeanos y wittgensteinianos.

Mayor fama, al menos en el ámbito de la filosofía de la mente, atesoran incluso los precedentes de la utilización de electrometeoros como término análogo para algún aspecto de la relación entre la mente y el cerebro: el propio Place (1956: 47) justificaba la imposibilidad de detectar rastros de la conciencia mediante la observación de la actividad cerebral –un *tópos* de insigne linaje: *cf*. las reflexiones de Leibniz en torno al molino pensante (Leibniz 1715: §17, *supra*)– apelando a que, también en el caso de los relámpagos, "[…] the operations for determining the occurrence of lightning are radically different from those involved in determining the occurrence of a motion of electric charges", sin que de ello deduzcamos que el relámpago y la descarga eléctrica sean realidades diferentes.

Sea como sea, ni los razonamientos de Stich ni los de Braddon-Mitchell y Jackson suenan al fin y al cabo demasiado diferentes de las terminantes objeciones con que Searle (1992) daba por liquidada la enmienda eliminacionista a la existencia de lo mental. La piedra angular sobre la que se yergue el eliminacionismo es un argumento –dice Searle (1992: 46)– "[…] so breathtakingly bad that I fear I must be misunderstanding it". Si hemos de resignarnos al convencimiento de que no existen pensamientos, deseos o alegrías, *porque* los conceptos teóricos de una neurociencia madura no puedan forzarse a encajar con esos conceptos de nuestra psicología natural –pensamiento, deseo, alegría–, entonces –señala Searle– habremos de resignarnos también a la inexistencia de muchas otras cosas: casi, en realidad, de todas las que conocemos. Así, *mutatis mutandis* –en una enumeración tan descabellada como la doctrina que se impugna:

> Physical theory covers the same domain as our commonsense theories of golf clubs, tennis rackets, Chevrolet station wagons, and split-level ranch houses. Furthermore, our ordinary folk physical concepts such as "golf club," "tennis racket," "Chevrolet station wagon," and "split-level ranch house" do not exactly, or even remotely, match the taxonomy of theoretical physics. [...]
>
> Therefore, split-level ranch houses, tennis rackets, golf clubs, Chevrolet station wagons, etc., do not really exist. (Searle 1992: 47)[27]

Si atendemos al argumento que subyace a los ejemplos aducidos por Stich (1992), la escasa viabilidad del proyecto de naturalización muestra perfiles muy semejantes a los que exhibiría en Horgan (1994, *supra*). La clave, en efecto, reside en el requisito de fijar condiciones necesarias y suficientes, descritas en un lenguaje naturalista, para la ocurrencia de tal o cual fenómeno intencional:

> Philosophical theories about mental representation typically offer what purport to be necessary and sufficient conditions for claims of the form:
>
> Mental state *M* has content *p*.
>
> And objections to these theories typically turn on intuitive counterexamples –cases in which the definition says that *M* has the content *p*, but intuition denies it, or vice versa. […T]here is now a fair amount of evidence suggesting that the assumptions underlying this traditional philosophical project may be simply mistaken. And if they are, then the project which dominates the philosophical literature on mental representation will be seriously undermined. (Stich 1992: 352)

[27] Aunque eso no constituya por fuerza un indicio de que Searle, efectivamente, haya malinterpretado el argumento eliminacionista, cabe señalar que, a diferencia de lo que ocurre con las creencias, los deseos, o las emociones, nadie había sugerido que las raquetas de tenis o los palos de golf sean el referente de conceptos teóricos: su carácter observacional, en el marco de la "física de sentido común", parece fuera de duda. Que éste adjudica equivocadamente a los conceptos psicológicos carácter teórico es la base de la crítica del eliminacionismo blandida por Heil 1989b: 349-350, *infra*.

La evidencia en cuestión, como en la cuarta premisa del argumento de Horgan (1994, *supra*) proviene de la propia investigación psicológica sobre formación de conceptos. Ahora bien, ¿qué conclusión se seguiría del hecho de que los conceptos empleados en la "práctica cotidiana" (Stich 1992: 351) de identificar los estados mentales según sus contenidos, en la cual se enraíza la psicología natural o de sentido común, respondieran más a modelos del proceso de categorización como los de Rosch, Barsalou o Rips que al modelo clásico de condiciones necesarias y suficientes? Si la conclusión no es la eliminación de los conceptos de la psicología del sentido común, ha de ser al menos, para Stich, el abandono del proyecto tradicional de construir una teoría de la representación mental mediante la definición de condiciones necesarias y suficientes, y su sustitución por el desarrollo de modelos cognitivos de aquellos procesos que involucren la idea de representación mental (Stich 1992: 354). En el argumento eliminacionista que parte del rechazo conjunto de que los conceptos psicológicos –o las clases que dichos conceptos definen– puedan identificarse con clases neurológicas ni con clases funcionales, sin embargo, parece intervenir subrepticiamente un dilema que nos fuerza a elegir entre expresar la extensión de un concepto por medio de condiciones necesarias y suficientes –sean éstas de uno u otro tenor, neurológicas o funcionales– y acatar la eliminación del concepto. Tal dilema, si Stich está en lo cierto, es falaz.

Otro argumento eliminacionista, por cierto, en el que Stich (1992) encuentra contratiempos parte de constatar la diferencia entre el concepto de representación mental empleado en las ciencias cognitivas y el de la psicología de sentido común –los dos proyectos de teoría de la representación mental que el propio Stich propone diferenciar. El argumento requeriría que el eliminacionista llegara a mostrar que "[…] there are some features that any scientifically respectable notion of mental representation will have to have, and [...] that these features are not endorsed by the account of mental representation implicit in folk psychology" (Stich 1992: 358). Los contratiempos vislumbrados por Stich tienen que ver con los presupuestos acerca de la determinación de la referencia de los términos teóricos que sería preciso aceptar si hemos de llegar hasta la eliminación partiendo de la divergencia entre los conceptos de representación mental de la psicología de sentido común y de las ciencias cognitivas.

> What the premises do entail is that folk psychology and computational theories make different and incompatible claims about the states they talk about. But that surely is not sufficient to show that they are talking about different things. If it were it would be all but impossible for theorists to disagree. (Stich 1992: 358)

Parece ser que Stich da por sentado, pese a sus críticas, que alcanzar la conclusión de que la psicología natural y la psicología cognitiva tratan sobre cosas distintas respaldaría la eliminación de los conceptos empleados en la psicología natural y la inexistencia de sus referentes. En todo caso, incluso para arribar al territorio intermedio –la psicología natural y la psicología cognitiva tratan sobre cosas

distintas– hace falta, según Stich, valerse implícitamente de una teoría descriptivista de la referencia, que permita pasar de la constatación de la diversidad de lo que predica sobre las representaciones mentales en uno y otro ámbito a la diversidad de aquello sobre el cual se afirman unos u otros predicados. Pero eso haría al eliminacionismo trivialmente verdadero:

> If minor disagreements between what commonsense says about mental states and what cognitive science says about them are sufficient to show that commonsense and cognitive science are positing different entities, then of course eliminativism is correct. But who cares? No one ever thought that commonsense psychology would turn out to be right about everything. (Stich 1992: 358)

Adoptar en cambio, una teoría histórico-causal de la referencia haría al eliminacionismo trivialmente falso: ningún concepto sería susceptible de eliminación; incluso del concepto de flogisto cabría decir que su referencia es el oxígeno, si bien el oxígeno tan sólo tiene alguna de las propiedades por medio de las cuales el concepto de *flogisto* determinaba su referencia. Stich (1992: 359) atribuye una tesis de este estilo a Lycan, quien se muestra:

> […] entirely willing to give up fairly large chunks of our commonsensical or platitudinous theory of belief or of desire (or of almost anything else) and decide that we were just wrong about a lot of things, without drawing the inference that we are no longer talking about belief or desire […]. I expect that "belief" will turn out to refer to some kind of information-bearing state of a sentient being, […] but the kind of state it refers to may have only a few of the properties usually attributed to beliefs by common sense. (Lycan 1988: 31-32 *apud* Stich 1992: 359)

La conclusión última de Stich es que el eliminacionismo "[…] makes no sense –it has no determinate truth conditions– unless it is tied to some specific account of reference" (Stich 1992: 360). Difícilmente podría el eliminacionismo, si esto es así, constituir *per se* una teoría –ni siquiera una concepción radicalmente deflacionista– de la intencionalidad de los estados mentales, puesto que depende para cobrar su propio sentido precisamente de eso: de una teoría del contenido de nuestros conceptos y, en general, de la intencionalidad de nuestros estados mentales. Así que –concluye Stich– el eliminacionismo no es ni mucho menos la amenaza que Fodor (1987: *xii*) o Dretske (1988: *x*) ven en él; probablemente ni siquiera sea una tesis particularmente interesante.

También Heil (1989b), por distintos caminos, ha hecho hincapié en el carácter autorrefutatorio de las tesis eliminacionistas –o abolicionistas, según la denominación que él emplea. Los motivos para rechazar dichas tesis son tan tajantes, a su juicio, como que de ser ciertas sería sencillamente imposible *aceptarlas* –o *rechazarlas*:

> [...I]t is idle to insist that any theory having as a consequence that we altogether lack contentful mental states is false a priori. To be sure, we may worry that such a theory, if

> true, could neither be taken seriously nor accepted: takings and acceptings would be mere fictions. And we swoon at the prospect of a theory that, if apt, must be simultaneously unbelievable and indubitable. A doctrine with these remarkable features seems at a minimum conceptually unstable. In setting out to abolish beliefs, it relinquishes its claim to be a theory we ought reasonably [to] believe. In the same way, we may be put hard to see how it could be thought to be *true*: Truth is precisely the sort of semantic doodad the theory bids us scorn.
>
> […]This is simply an intriguing, if kinky, consequence of the theory. It cannot be believed or asserted –it cannot even be true– in any world that satisfies it. (Heil 1989b: 346-347)

Sería prematuro, sin embargo, despachar el eliminacionismo con la displicencia que el argumento de Heil parecería a simple vista requerir. Particularmente elegante resulta la réplica que, pese a su rotundo rechazo de la concepción eliminacionista de las relaciones entre lo psicológico y lo físico, esgrimen en su defensa Braddon-Mitchell y Jackson (1996):

> Eliminativists must concede that until we have to hand the new categories that will, they hold, eventually come to replace belief and desire in the project of explaining behaviour, we perforce must talk in the old way. But, they can argue, when the category that comes to replace belief –say, belief*– is to hand, and the category that comes to replace desire –say, desire*– is to hand, they can say, without paradox, what their eliminativist view is: they believe* that there are no beliefs and desires, and desire* to tell us this. (Braddon-Mitchell y Jackson 1996: 242)

Hay, no obstante, un flanco que la táctica dialéctica de Braddon-Mitchell y Jackson deja desguarnecido, y por el cual la refutación del eliminacionismo ensayada por Heil puede recobrar su fuerza. El problema reside en que no nos es dado suponer que el eliminacionismo sea compatible con la existencia de categorías –*creencia**, *deseo**– que, habiendo quedado delimitadas según criterios no propiamente psicológicos –fisiológicos, por ejemplo– "vengan a reemplazar" a nuestras categorías ordinarias de *creencia* o *deseo*, es decir, que resulten en alguna medida razonable coextensivas con éstas. Tal suposición resulta inadmisible porque, precisamente, si existieran tales categorías, cabe esperar que podríamos formular algún mecanismo de reducción interteórica para traducir las explicaciones psicológicas que apelaran a creencias y deseos a explicaciones fisiológicas que apelaran en su lugar a creencias* y deseos*. Y que eso suceda es tanto como que el eliminacionismo quede desacreditado en favor de una u otra variante de reduccionismo. Dicho de otro modo: lo que el eliminacionista vaticina es justo que las categorías de la psicología cotidiana, como *creencia* y *deseo*, *no encontrarán reflejo* en una ciencia madura de la conducta, o del funcionamiento del sistema nervioso. Con palabras prestadas de P.M. Churchland, uno de los más ardientes defensores de la necesidad de arrojar el "marco de

referencia psicológico", siguiendo los pasos de Bernardino de Siena o Savonarola, al *falò della vanitá*[28]:

> A juicio del materialismo eliminativo, no podrán encontrarse las correspondencias biunívocas [entre los conceptos de la psicología corriente y los conceptos de la neurociencia teórica], y no se podrá efectuar una reducción interteórica del marco de referencia psicológico corriente, *porque el marco de referencia psicológico que utilizamos corrientemente es una concepción falsa y radicalmente engañosa sobre las causas de la conducta humana y la naturaleza de la actividad cognitiva*. (Churchland 1984/1988: 75)

Así pues, ni siquiera usando el vocabulario teórico de esa ciencia madura podrá expresarse la generalización según la cual los eliminacionistas *creen* que no existen creencias –generalización que el eliminacionista debe, de hecho, dar por falsa: es de suponer que no hay *nada significativo* en común, desde su punto de vista, entre los sujetos a los que atribuimos, a tenor de nuestra desatinada teoría informal sobre los determinantes internos de la conducta, tal o cual creencia.

En todo caso, un leve giro en el argumento de Heil puede agravar los estragos que ocasiona sobre la posición eliminacionista. Se trata de reclamar al eliminacionismo no ya una reconstrucción de la evidencia de que los propios eliminacionistas albergan creencias –evidencia que a fin de cuentas está en su mano negar–, sino de la de que mantengan *teorías* –ellos, o cualquier otro. En ausencia de una reconstrucción de la idea de *mantener una teoría* en la que se acredite la expurgación de todo residuo mentalista –es decir, en la que mantener una teoría no acabe implicando *albergar ciertas creencias*–, se vuelve contra el proyecto de abolición del discurso psicológico el hecho de que dicho proyecto descansa sobre la tesis de que los conceptos que integran el discurso psicológico –originalmente, los conceptos mentalistas de la psicología cotidiana: creencia, deseo, etc.– forman parte de una *teoría* que mantenemos tácitamente desde tiempo inmemorial, y que empleamos, fallidamente, para comprender la conducta de los demás y la nuestra propia. Si se desprovee al eliminacionista de la premisa según la cual los conceptos mentalistas de

---

[28] Acaso la comparación no se nos antoje, salvando las distancias, *del todo* gratuita al reparar en que, como el confesor de Lorenzo de Médici, Churchland se deleita en profetizar un nuevo tiempo, si bien no será a su juicio un renacido Ciro quien ordene a sangre y fuego las impías costumbres de los toscanos, sino una nueva ciencia materialista quien ilumine los recovecos de nuestra comprensión de nosotros mismos:

> […] la genuina llegada de una cinemática y una dinámica materialistas para estados psicológicos y procesos cognitivos […] constituirá, no una penumbra en la que nuestra vida interior será eclipsada o suprimida, sino un amanecer en el que sus maravillosas complejidades serán finalmente reveladas –tranquilamente, si nos concentramos en la introspección consciente. (Churchland 1984/1988: 256)

Esta forma de asumir sin titubeos todas las consecuencias de una hipotética eliminación del vocabulario psicológico cotidiano, no sólo de sus reliquias en la teorización científica, sino también de su uso corriente, es sin duda el aspecto más llamativo de la propuesta eliminacionista de Churchland.

la psicología cotidiana forman parte de una teoría, entonces queda del todo inerme la premisa según la cual los términos teóricos propios de las teorías erróneas deben quedar relegados a la apretada historia de nuestra ignorancia a la vez que se reconoce la inexistencia de sus referentes. En cambio, si la reconstrucción eliminacionista del concepto de mantener una teoría prospera sin el recurso velado a otros estados mentales, tendremos que al menos un concepto mentalista –el de mantener una teoría– ha quedado adecuadamente reconstruido en términos ajenos al vocabulario de la psicología cotidiana, y merece, por tanto, mejor destino que la eliminación –su reducción ha quedado reivindicada. Esto es, en suma: mientras el eliminacionismo no proporcione una depuración de la idea de mantener una teoría, no proporciona tampoco motivos para aceptar que creencias y deseos puedan correr la misma suerte, con el tiempo, que aguardaba al flogisto, los espíritus animales, el mesmerismo, la posesión demoníaca, el éter, o el aborrecimiento del vacío, pero si el eliminacionismo proporciona tal depuración, se inflinge al hacerlo su propia derrota.

Esta inclinación autodestructiva de la doctrina eliminacionista arraiga en realidad –piensa Heil– en cierta incapacidad de distinguir *explanandum* y *explanans*:

> It is important, surely, to distinguish, at least at the outset, *phenomena* that we wish to explain from specialized entities postulated in explanations. […I]ntentional properties appear to occupy space, as it were, in the empirical world. Their status resembles that of tables and trees, items requiring accomodation, not hidden esences. We shall continue to need an account of intentional phenomena even if cognitive science, for reasons of its own, should elect to exclude reference to meaningful states and processes from theories of cognition. Eventually […], we shall insist on theories that incorporate accounts of our ability to think and talk *about* things. No psychology that failed to offer such an account could be considered complete. (Heil 1989b: 349-350)

O, más taxativamente:

> […I]ntentional goings-on make up important parts of what we wish to understand about ourselves and our world. If our categories fail to encompass these, so much the worse for our categories. The task is to explain intentionality, not to explain it away. (Heil 1989b: 364)

El origen de la táctica ensayada por Heil ante el asedio eliminacionista estaría –si Lycan (1986b) está en lo cierto– en las reflexiones de Sellars (1956, 1963),

> […] for as Sellars originally saw, public (particularly linguistic) behavior still precedes intentional psychology in the order of explanation; the notion of public speech is prior to that of private speech or thought and is only later explained by it. In real life, public speech episodes serve as data, in their semantical *rather than* their brute-behavioral guises; people learn semantical descriptions of verbal episodes long before they learn brute-behavioral descriptions, if they *ever* learn brute-behavioral descriptions of oral speech. (Lycan 1986b: 177)

El propio Lycan señala con acierto el paralelismo entre la argumentación sellarsiana en este punto y la crítica común, casi canónica, de la noción de *dato sensorial* –una de cuyas fuentes primigenias es, por supuesto, el feroz ataque del propio Sellars contra el mito de lo Dado. Efectivamente, el propósito de relegar a la descripción intencional y teleológica de la conducta del papel de dato de la explicación psicológica, en beneficio de una pretendida descripción cinética, guarda cierta semejanza con el malhadado proyecto de reemplazar las descripciones ordinarias de nuestro entorno perceptivo por la enrevesada jerigonza de los datos sensoriales. Pero si lo que la psicología de la percepción debe explicarnos es por qué nos parecen dulces las manzanas, y no porque categorizamos como dulzor los estados perceptivos de naturaleza gustativa y somatosensorial que elicitan en nosotros las manzanas, entonces la psicología de la conducta deberá explicarnos por qué intentamos alcanzar la manzana madura del árbol, y no por qué uno de nuestros tríceps braquiales provoca la extensión de las articulaciones húmero-radial y húmero-cubital, levantándose así nuestro brazo en determinado ángulo y con determinado patrón de aceleración y deceleración, a la vez que se produce una hiperextensión de las articulaciones rotulianas, etc. La cuestión de qué vocabulario es el apropiado para la descripción de estímulos y respuestas acabaría por convertirse, como tendremos ocasión de analizar pausadamente, en uno de los ejes fundamentales que modulan tanto la ruptura entre la concepción conductista de la psicología y el cognitivismo, como la continuidad entre ambas escuelas.

Conviene dejar apuntado, por último, que de manera más nítida que en las críticas de Sellars (1956) acaso sea en Carnap (1956: 70, *supra*), en la propia órbita de lo que Sellars denuncia, donde quepa rastrear las fuentes de las que sin duda mana la confusión entre *explanandum* y *explanans* que Heil quiere impugnar: de la premisa de que los conceptos mentalistas son de naturaleza teórica es fácil deslizarse hasta la conclusión de que no forman parte del *explanandum* de una ciencia madura de la conducta, sino del *explanans* de sus toscas predecesoras.

A pesar a lo notoriamente difícil que resulta, como hemos visto, hacer del eliminacionismo siquiera una posición coherente, el tentador influjo de la contienda entre lo inexplicable y lo inexistente sigue, con todo, dejándose sentir en el discurso científico y filosófico. Así, puede leerse que:

> En realidad, no existe tal cosa como "la mente". Es una idea que propagamos: que hay una cosa que se llama "la mente". Personalmente me gusta la definición de Marvin Minsky de la mente como "lo que hacen los cerebros". La mente es eso, en cuyo caso "la mente" en realidad no es nada. (Blackmore 2008)

Conviene asumir que las palabras de Blackmore no han de tomarse al pie de la letra, pues de lo contrario un mero ejercicio de transitividad nos abocaría a concluir que "lo que hacen los cerebros […] en realidad no es nada" –hicieron falta trabajos experimentales como los de Gustav Fritsch y Eduard Hitzig (1870) para deshacernos de la antigua doctrina de la inexcitabilidad de la corteza cerebral, pero ni siquiera

existe una antigua doctrina de la *inactividad* del córtex. Ahora bien: si no es en sentido literal, no es fácil averiguar en qué sentido debe tomarse la aseveración de que "no existe tal cosa como 'la mente'". Tal vez, desde luego, sólo quiera decir que no todas las propiedades que venimos atribuyendo a "la mente" se verán ratificadas por la investigación científica, o que no todas ellas formarán parte del vobulario teórico de la ciencia, pero esto, como hemos visto, puede predicarse no sólo de "la mente", sino de prácticamente cualquiera de las cosas que nos rodean.

Al lado de la alucinación de lo inexistente, con todo, sigue en pie la fascinación de lo inexplicable:

> […S]i aceptamos una visión del mundo ordinaria, realista, existe la materia y la cosa física; puedes decir que nació un bebé y que creció y sigue viviendo […] Si entonces giras hacia el otro lado y preguntas: ¿Qué es todo esto desde su propio punto de vista? Esto es un misterio, y personalmente no lo sé. (Blackmore 2008)

Así, también, junto a la rotunda declaración de que "[…] francamente no hay nada que no sea este cuerpo moviéndose y haciendo cosas", encontramos una confesión no menos franca: "[…]¡[M]ira ese color! ¿Cómo puede ser eso actividad cerebral" –"[…] De alguna manera, de alguna manera, pero ¡no veo cómo! ¡No lo comprendo!" (Blackmore 2008).

A fin de cuentas –qué duda cabe– es comprensible que esta tensión acabe dando fuerza al vivo aliento poético que puede albergarse en el árido seno de un trabajo sobre codificación neuronal en los sistemas de control psicomotriz de la rata blanca de laboratorio (*Rattus norvegicus albino*); acaso no resulte del todo inesperado, al menos para el lector atento de Place (1956: 47, *supra*) y Braddon-Mitchell y Jackson (1996: 253-254, *supra*), que el término figurado de la metáfora que anima dicho aliento poético sea precisamente una tormenta estival:

> Nuestras facultades de hablar, de amar, de odiar y percibir el mundo que nos rodea, así como nuestros recuerdos, nuestros sueños e incluso la historia de nuestra especie, emergen de la combinación de una multitud de diminutas señales eléctricas que se difunden por el cerebro, lo mismo que una tormenta eléctrica llena de relámpagos el cielo de una noche de verano. (Nicolelis y Ribeiro 2009: 10)

Coda. Quimera de la nube equivocada: la naturalización y el error

Pues bien: en medio de las enormes dificultades que, como anotan Domínguez (2003: 42) o Moya (1994: 233), *supra*, plagan la ardua tarea de urdir una explicación naturalizada del don en virtud del cual nuestros pensamientos, o nuestras palabras –*phoné semantiké*, dejó dicho Aristóteles: *cf*. *Poética* 1457*a*; *Política* 1253*a*–, se ligan a las cosas mundanas, no tardaría en hacerse patente que lo que habría de exigírsenos entender no es sólo el hecho de que pensamientos o palabras puedan designar o describir el mundo, sino también el de que puedan hacerlo equivocadamente –sin

dejar, por ello, de ser pensamientos, o palabras. En idéntico trance, como hemos advertido, acabó viéndose Frege al tratar de desentrañar la noción de sentido que se desprendía de sus investigaciones sobre la relación de identidad expresada por el símbolo "=". La cuestión del error ha ido cobrando un creciente relieve en los intentos de naturalización de las propiedades semánticas e intencionales, hasta llegar a perfilarse a veces como vertebradora del propio proyecto:

> How do thoughts, purely natural, physical states of persons, get to be about the world and acquire their semantic properties? Call the project of answering these questions the "naturalization project". The "naturalized" part of the project is to explain how something with meaning can arise out of purely natural causes and physical objects without depending upon things that already have meaning. It is, if you will, the project of explaining how there can be underived intentionality –unmeant meaners. The "semantic" part is the attempt to get beyond the level of information or indication to the level of meaning, to the level where representation and falsehood are possible. (Adams y Aizawa 1994: 223)

Es esto es lo que subyace a la expeditiva identificación de la semántica como "el nivel donde la representación y el error son posibles" que plantean Adams y Aizawa (1994, *supra*): el convencimiento de que la relación entre un pensamiento –o una oración– y el estado de cosas al que se refiere tiene un sustrato natural en lazos tan ubicuos como los que unen a los anillos del interior del tronco de los árboles con las condiciones climáticas en que crecieron –de modo que las técnicas dendrocronológicas hacen viable una estimación de la edad del árbol–, o las que unen la presencia de cúmulonimbos con la probabilidad de tormentas, salvo porque en ese nivel de mera "información o indicación" natural, como hemos visto, falta la posibilidad del error, que se convierte así en el penúltimo grial de la indagación acerca de la naturaleza de lo mental[29].

La existencia de signos naturales, como las nubes que amenazan lluvia, ha constituido reiteradamente un punto de partida de la investigación en torno a la naturaleza de los signos lingüísticos, o de sus trasuntos mentales. Pero ninguna nube –ya lo sabemos– indica erróneamente que se acerca un aguacero, por lo mismo que ninguna nube representa erróneamente la figura de un caballo o un ángel. Según el relato de su *Vida* que en los albores del s. III nos legó el neopitagórico Flavio Filóstrato, Apolonio de Tiana se habría esforzado en convencernos de que las

[29] Lo que caracteriza, así pues, al ámbito del significado que se abre con lo mental es la posibilidad del error –no, *nota bene*, la materialización de esa posibilidad. Tiene razón Mackie cuando apunta, al hilo del intento de perfilar la concepción de la mente propugnada por Locke como una teoría representacional en la que las ideas lockeanas se entiendan como objetos intencionales, que:

> Even if there were no illusions, no perceptual errors, there would still be a difference between the table's being square and my seeing it as square: how I see it would be distinct from how it is, even if they were exactly alike, even if I always saw it exactly as it is. This distinctness is established by the mere logical possibility of perceptual errors and distortions, but it is made doubly secure by their actual occurrence. (Mackie1976: 48-49)

mudables formas que remedan las nubes son fruto del mero azar, no de la mano caprichosa de los dioses. Vemos, entonces, animales o monstruos en ellas, pero no están en ellas como lo estarían si a alguna divinidad se le hubiera antojado. Como no están en ellas, no pueden estar *erróneamente* en ellas[30].

En nuestros pensamientos, en cambio, y de modo particularmente patente en nuestras palabras, sucede algo muy distinto. Como tantas veces se ha dicho, "[…e]l carácter intencional del conocimiento se revela primariamente en el lenguaje, en la evidencia irreductible de que nuestras palabras significan algo" (Llano 1999: 107). Esa evidencia irreductible, sin embargo, no nos parece hoy tan transparente como pudiera parecérselo al hijo e interlocutor de Agustín de Hipona –al contrario, no es difícil que convengamos con Ogden y Richards (1923: 10) en verla como el "centro de la obscuridad tanto en teoría del conocimiento como en cualquier otra discusión":

> *Agustín*: Me gustaría saber cómo replicarías a uno del que solemos reírnos al oír que llegó a esta conclusión: que había salido un león de la boca de aquel con el que estaba discutiendo. Efectivamente, le había preguntado si las cosas que decimos salen de nuestra boca, y el otro no lo pudo negar. Después, le fue fácil lograr que su hombre, en el curso de la conversación, nombrara al león. Y, tan pronto lo consiguió, comenzó a insultarlo ridículamente insistiéndole en que, como él había confesado que todo lo que hablamos sale de nuestra boca, y acababa de pronunciar 'león', no podía negar que un hombre nada malo parecía haber vomitado una bestia tan enorme.
> *Adeodato*: Claro que no era nada difícil enfrentarse a ese burlón, puesto que yo no le concedería que sale de nuestra boca todo lo que hablamos. Porque lo que hablamos, lo significamos; pero de la boca del que habla no procede la cosa que es significada, sino el signo con que se la significa. (Agustín, *El Maestro* VIII, 23; 2003: 102-103)

Pretender que quien habla exhale la cosa significada y no su signo era una broma filosófica común al menos desde el tiempo de la segunda Stoa. Como bien ha recordado Sorensen (2003: 44), en efecto, Diógenes Laercio atribuía a Crisipo el "argumentillo" siguiente: "Si dices algo, ello pasa por tu boca; *atqui*, dices carro,

[30] Aunque puede ocurrir, también, que el hallazgo que nos depara la naturaleza ejerza tanta violencia sobre nuestra concepción del mundo que no podamos sino tomar por enteramente azaroso, como formas en la nubes, lo que en realidad no lo es. Así, sabemos que entre las hipótesis que manejó la Antigüedad para tratar de entender los fósiles de los que aquí o allá se iba teniendo noticia no resultaba descabellada la de que se tratara de meros divertimentos, inocentemente provocados por fuerzas naturales al dotar a una mera piedra de un asombroso parecido con algún animal que acaso nunca hubiera existido. Como apunta Pimentel (2010: 208) en su espléndido estudio sobre los vínculos entre la historia de dos descubrimientos, el del rinoceronte y el del megaterio:

> Como las copias o las imitaciones, los fósiles eran duplicados, correspondencias, *imágenes* de otros seres vivos, una cuestión con ecos en la caverna platónica […] ¿Lo eran? ¿O simplemente se parecían? El tema de las armonías y las resonancias ocultas de los objetos naturales está cruzado por las (problemáticas) relaciones entre arte y naturaleza […]: la metafóra del mundo como un artificio es una idea que atraviesa la ciencia, la poesía, el teatro y la pintura de la Edad Moderna […]. Desde este punto de vista, los fósiles eran *lusus naturae*, artificios naturales, entrenamientos o juegos de la Naturaleza, demostraciones de sus virtudes plásticas. Al igual que el hombre imitaba al Creador con el arte y la ciencia […], no resultaba extraño que la propia Naturaleza imitara mediante estos ingenios su propia obra, la *reprodujera*.

luego un carro pasa por tu boca" (*Vidas* VII: 7). Los lazos entre las palabras y las cosas –*onósmata* y *prágmata*, en los términos que dejaría acuñados Clemente de Alejandría– conformaban aún para Agustín una "[...] ley de la razón, impresa en nuestras almas [...]", una "regla que por naturaleza es la más poderosa" en virtud de la cual una vez oídos los signos, la atención se dirige a las cosas por ellos significadas" (Agustín, *El Maestro* VIII, 24; 2003: 103-104). La *suppositio formalis*, pues, se impone, como sentido primario y natural de las palabras, por encima de la *suppositio materialis*: mencionar una palabra es el acto reflexivo que sólo surge una vez que ésta ha recibido su uso. Todo esto, desde luego, es como debe ser, pero la naturalidad de que se reviste el uso del signo por la cosa significada puede tal vez hacernos desatender la perplejidad que ese hecho natural despierta: que la realidad material de los signos pueda atesorar la cualidad de estar por las cosas significadas –o, si se prefiere, que dichas cosas puedan existir intencionalmente en los signos– no es por más natural menos fascinante, y, por supuesto, no queda explicado al hacer ver su naturalidad[31]. Ahora bien, que cualquier avance en nuestra comprensión del significado nos acerca quizá como ninguna otra cosa a entender lo mental es algo que ya parece haber visto con claridad un pionero de la investigación sobre la memoria como T.H. Pear (1923: 59 *apud* Ogden y Richards 1923: 12), según revela su rotunda afirmación respecto a que "[…s]i el descubrimiento de la naturaleza psicológica del Significado tuviera un éxito completo, podría constituir un fin definitivo para la psicología".[32]

Es indudable, así pues, que la relación entre el problema mente-mundo y el problema mente-cuerpo era más nítida para Agustín –como lo había sido para los griegos– de lo que es hoy para nosotros:

> Puesto que el mismo nombre consta de sonido y de significado, y el sonido pertenece al oído y el significado a la mente, ¿no consideras que en el nombre, como en un ser animado, el sonido es el cuerpo y el significado, en cambio, es como el alma? (Agustín, *De la cantidad del alma* XXII, 65; 2003: 147)

El descubrimiento agustiniano del *sermo interior* –"Pues aunque las palabras no suenen, quien piensa las dice ciertamente en su corazón", dice el tagastino en *De la Trinidad* XV, 10, 17 (2003: 162)–, el alumbramiento de ese flujo de conciencia que sería mucho después la divisa de William James (1890, 1892), se perfila así a la par como tempranísima descripción de la *lingua mentis* de la ortodoxia cognitivista: un lenguaje entendido como anterior al lenguaje, que es sustrato de éste y del que éste es signo.

---

[31] La observación es la misma que, al hilo de las palabras de Bechtel (1988: 73) y Rodríguez (2001b: 225), se hacía valer *supra* acerca de la noción de actitud proposicional tal como se emplea desde Russell (1940).

[32] La cita de Pear por parte de Ogden y Richards es más bien una reconstrucción, fiel aunque inexacta, que parece provenir de estas líneas:

> […] if this discovery [the discovery of the *psychological* (as distinguished from the logical or the metaphysical nature of meaning)] were completely succesful it would put an end not only to the problem of meaning, but to psychology altogether, for every psychological investigation deals with some aspect of this question. (Pear 1923: 59)

Todo encaja, una vez más, en la esforzada cifra de las investigaciones de los antiguos a la luz de su propia fe que Agustín, desde su renuncia al maniqueísmo, dedicara tantas horas a labrar:

> Por consiguiente, la palabra que suena fuera es signo de la palabra que brilla dentro, a la cual responde mejor el nombre de 'palabra'. [...] En efecto, nuestra palabra se hace de algún modo voz del cuerpo, en cuanto que asume aquella en la que se manifieste a los sentidos de los hombres, del mismo modo que el Verbo de Dios se hizo carne, en cuanto que asumió aquella en la que también él se manifestara a los sentidos de los hombres. (Agustín, *De la Trinidad* XV, 11, 20; 2003: 162-163)

De esa manera, la incipiente consciencia del hiato entre palabras y cosas que había ido ahormando el pensamiento griego comienza a impregnar una visión del mundo, la cristiana, que se había venido mostrando ajena a la distancia entre el nombre y lo nombrado. No en vano relata el *Apocalipsis* (11: 13), como nos recuerdan Ogden y Richards (1923: 52), que "[…m]urieron en el terremoto siete mil nombres de hombres"; en la misma línea anota Laín Entralgo (1987: 33) algunas conclusiones de Tresmontant (1953: 100) acerca de las líneas maestras de la metafísica bíblica, que queda descrita como "una metafísica del nombre, del nombre propio", apuntando al paso el papel primordial que el *Génesis* concede a la palabra creadora, y explicando así las diferencias entre la magia nominal de los pueblos semíticos y la *epaoidé* –el ensalmo– con que los médicos griegos aspiraban no a dominar la naturaleza al nombrarla por su nombre verdadero, sino a complacer el corazón de los dioses: "el personalismo de la mentalidad semítica y el naturalismo de la mentalidad griega" (Laín Entralgo 1987: 35) se apartarían aquí, como en tantas otras cosas.

En efecto, cuando Flavio Filóstrato insiste en hacernos ver que las nubes no figuran tal o cual bestia más que al albur de innumerables coincidencias, atribuirle la idea de que, por tanto, nunca podrán figurarla equivocadamente sólo puede hacerse, sin forzar injustificadamente la interpretación de sus reflexiones, justo en la medida en que las fisuras entre nuestras figuraciones de las cosas –palabras, sobre todo, pero también, acaso más de forma más patente, representaciones plásticas– y las propias cosas han sido ya entonces escrutadas con cierta minuciosidad por el pensamiento griego. Que Heráclito (D-K 22 B1; Sexto Empírico, *Contra los profesores* 7.132) pudiera asegurar que "[…] aunque todas las cosas acontecen según este *Lógos* […]", "[…] siempre se quedan los hombres sin comprender que el *Lógos* es así como yo lo describo […]" es sólo uno de los muchos signos de que, en el seno de la comprensión de que el saber es un emparejamiento entre el orden de las cosas y el orden de las palabras o los pensamientos, comienza a discernirse poco a poco la extrema fragilidad de tal emparejamiento[33]. Al tiempo, entonces, que en la noción de *Lógos*

[33] De la multiplicidad de sentidos en que se habla del *Lógos* da fe el inspirador inventario que nos proporcionan Gallero y López (2009, *eds*.) en sus notas al citado fragmento D-K 22 B1; se hace transparente así que dichos sentidos apuntan con igual naturalidad al orden del mundo y al del pensamiento o la palabra. Así puede *Lógos*, en efecto, decirse como "[…] modelador de las cosas a través del movimiento de contrarios (Aecio, *Opiniones de los filósofos* I, 7, 22), como "[…] unidad de la

cristaliza esta íntima cohesión entre el mundo y el pensamiento, se ha develado su quiebra: en la indagación de los presocráticos, y en Heráclito particularmente, la propia existencia del error –no de la ignorancia indolente, sino del pensamiento equivocado o de la palabra engañosa– comienza a adivinarse como asunto de esa indagación –o, si se admite el anacronismo, como *quaestio disputata*.

Esta fractura que entrevemos en Heráclito es apenas uno de los muchos presagios –la insistencia de Parménides (D-K 28 B1; Sexto Empírico, *Contra los profesores* 7.132 y Simplicio, *De caelo* 557, 25) en distinguir "[…] entre el imperturbable corazón de la Verdad persuasiva" y "[…] las opiniones de los mortales, en las que no hay verdadera creencia" constituye un ejemplo cuando menos igual de rotundo que el de Heráclito[34]; tal vez otro tanto pueda decirse del propio carácter aporético del pensamiento de Zenón de Elea, o al menos de las paradojas que forman lo que de él conservamos– por medio de los cuales se anuncia la investigación acerca de la capacidad de las palabras para nombrar erróneamente que atraviesa y da sentido al coloquio etimológico, tan arduo como disparatado, que entre frecuentes guiños irónicos arrostra Sócrates en el *Crátilo* platónico. El punto de partida del diálogo atañe expresamente a la posibilidad de que las palabras –los nombres propios, en particular– yerren en su concomitancia con las cosas: Crátilo mantendrá que la exactitud pertenece a la naturaleza (*katà phýsin*) misma de los nombres, y se mostrará dispuesto a extender su tesis a toda palabra (*Crátilo* 431*b-c*), mientras Hermógenes le

---

contraposición" (Hegel 1816), "[…] armonía oculta de las fuerzas opuestas, […] unidad profunda que ocultan y traducen las aparentes oposiciones" (Schuhl 1934), "[…] ensamblaje original, carácter constitutivo del Ser mismo" (Heidegger 1944), "[…] fundamento del mundo, […] norma que todo lo determina, […] ley cósmica" (Fränkel 1951), "[…] relación de una cosa con otra, conjunción, […] armonía del conjunto (Heidegger 1951/1953), "[…] sustrato unificante de la pluralidad, […] *coincidentia oppositorum*" (Marcovich 1968). Pero con igual propiedad puede *Lógos* venir a ser "[…] el acto de hacer sensible el pensamiento a través de los nombres y los verbos, de suerte que se refleje en la palabra como en un espejo o sobre el agua" (*Teeteto* 206c-d), "[…] la necesidad del pensamiento, la ley lógica […]" (Reinhardt 1916), "[…] un conocimiento del cual se originan, al mismo tiempo, la palabra y la acción" (Jaeger 1933), "[…] significado de la doctrina aquí presentada" (Diels y Kranz 1951), o sencillamente "[…] pensamiento expresado por palabras" (Gomperz 1953), "[…] todo lo que se dice de palabra o por escrito, incluyendo respuesta de un oráculo y expresión proverbial; valoración o estima; conversación con uno mismo o pensamiento; causa; verdad; medida; proporción; principio, norma o sentencia; facultad de la razón; definición o fórmula" (Guthrie 1962), "[…] sentido" (Brun 1965), "[…] *lo que se comunica*" (Gadamer 1978), "expresión" (Colli 1980: 25), o sencillamente el "habla" (Pinnola 2007). Incluso, afanándonos en verter en una única expresión ese recorrido entre el pensamiento –o el lenguaje– y su objeto, cabría hablar, con las precauciones que apuntan Kirk, Raven y Schofield (1957: 274), de una "[…] fórmula unificadora o método proporcionado de disposición de las cosas, […] su plan estructural", o más nítidamente, en la estela de Laín Entralgo (1958), entender el *Lógos* como "[…] aquello por lo cual el hombre puede dar razón de la realidad y expresarla […], una razón u orden inmanente".

Todas las referencias de esta nota, salvo Colli (1980: 25) y Kirk, Raven y Schofield (1957: 274) provienen de Gallero y López (2009, eds.: 239-245).

[34] En Parménides hemos encontrado ya *supra*, además, los indicios de un primer escrutinio de las dificultades lógicas que conlleva la mera existencia de pensamientos falsos, dificultades cuya plena conciencia engendraría buena parte del trabajo de Frege.

opondrá la convicción de que dicha concomitancia –y por tanto su exactitud o la falta de ella– son fruto de la convención (*katà nómon*) o el hábito (*katà éthos*); Sócrates les hará ver a ambos, uno tras otro y a veces casi entre burlas, que cualquiera de esas dos concepciones del origen del vínculo entre la palabra y la cosa nos forzaría a admitir que no "se puede hablar falsamente", cuando es obvio que "[…] lo mismo que un retrato se puede aplicar a quien no le corresponde, así el nombre puede aplicarse al objeto que no le corresponde […]" (Calvo 1983: 347; *Crátilo* 431*c-d*).

Evocando sin apenas veladura alguna palabras de Parménides, una de las primeras tesis que Sócrates trata de dejar establecidas es que cabe hablar falsamente, que "[…] es posible designar mediante el discurso a lo que es y a lo que no es […]" (*Crátilo* 385*b*)[35]. Hermógenes no puede dejar de concedérselo; Crátilo, en cambio, se opondrá vanamente a la convicción de Sócrates apelando a un razonamiento –no es posible que "quien dice lo que dice no diga lo que es"– que Sócrates, burlón, tilda de "un tanto sutil para mí y para mi edad" (*Crátilo* 429*d-e*). Lo que entienden Sócrates y sus interlocutores por exactitud de los nombres se expresa quizá en *Crátilo* 424*a* con mayor claridad que en otros fragmentos: "[…] si es verdad o no que captan el ser por medio de sílabas y letras hasta el punto de imitar su esencia"[36]; tal concepto del lenguaje –"que las cosas hayan de revelarse mediante letras y sílabas" es taxativamente rechazada por Sócrates, que lo tilda de "manifiestamente ridículo" (*Crátilo* 425*d*).

Aunque –qué duda cabe– sumamente cambiado por el tiempo, es éste el mismo atolladero al que Fodor (1984) intenta abocar a las teorías causales de la representación, como la de Dretske (1981) –un atolladero al que, de hecho, llama Fodor "el problema de Platón"[37]:

> […] causal theories have trouble distinguishing the conditions for representation from the conditions for truth. This trouble is intrinsic; the conditions that causal theories impose on representation are such that, when they're satisfied, *mis*representation cannot, by that very fact, occur. Hence, causal theories about how propositional attitudes represent have Plato's problem to face: how is false belief possible? (Fodor 1984: 3)

[35] Respecto de las creencias y los conocimientos hace lo propio Sócrates en *Gorgias* 454*d*, en el seno de la reflexión sobre la naturaleza y objeto de la retórica que da cuerpo al diálogo.

[36] *Cf*. también *Crátilo* 435*e*: si los nombres son inherentemente exactos en el sentido que Sócrates discute, "[…] cuando alguien conoce qué es el nombre […] conocerá también la cosa, puesto que es semejante al nombre".

[37] La expresión "el problema de Platón" suele emplearse en los ámbitos del cognitivismo para referirse al modo en que Chomsky, por medio de la noción de pobreza estimular, enlaza su defensa de la existencia de una gramática universal innata con el conocido fragmento de *Menón* (82*b*-86*a*) en el que Sócrates interroga a un niño iletrado –un esclavo del propio Menón– hasta extraer de sus respuestas el conocimiento del teorema de Pitágoras. Aunque Fodor (1984) no hace ninguna referencia al *Crátilo*, resulta obvio que el "problema de Platón"que plantea es más cercano al que Sócrates aborda con Crátilo que al que discutiera con Menón cuando éste se hospedaba en casa del magnate Ánito, que luego se contaría entre sus acusadores.

En realidad, claro, hacia lo que una teoría meramente causal de la referencia parece precipitarse es hacia el naturalismo de Crátilo: que la exactitud es inherente a los nombres *katà phýsin*. Es un lugar común de nuestra reflexión sobre el pensamiento y el lenguaje que representación, o referencia, son –como significado e intencionalidad– nociones normativas, y la de causalidad, que no lo es, parece inerme –como las de correlación e información– a la hora de articular la reconstrucción de la normatividad de las primeras. O, con palabras de Fodor:

> Information is correlation and though correlations can be better or worse –more or less reliable– there is no sense to the notion of a *mis*correlation: hence there is nothing, so far, to build the notion of misrepresentation out of. (Fodor 1984: 8-9)

El caso es que algo antes, en *Crátilo* 387*d*, Sócrates deja afianzada la distinción entre nombrar e intentarlo vanamente –no llegar siquiera a nombrar porque no se usan los nombres de la forma correcta, como no se llega a cortar, taladrar o tejer si no se hace debidamente: "[…] habrá que nombrar como es natural que las cosas nombren y sean nombradas y con su instrumento natural, y no como nosotros queramos […]"; "[…] en tal caso tendremos éxito y nombraremos, y, en caso contrario, no […]"[38]. Resta entonces descartar que todo nombrar falsamente sea un no nombrar[39]. A ello dedica Sócrates varios argumentos: en 432*e* lo escuchamos convencer a Crátilo de que incluso si una letra "[…] que no le corresponde […]" aparece en el nombre, o un nombre dentro de la frase, o una frase en el discurso, "[…] no por ello, deja de nombrarse o decirse la cosa, con tal que subsista el bosquejo de la cosa sobre la que versa la frase […]"; justo antes, en 432*d*, se aduce también que de darse la implacable correspondencia entre lenguaje y realidad que Crátilo imagina, "[…] todo sería doble y nadie sería capaz de distinguir cuál es la cosa y cual el nombre"; a continuación se despliega una cadena de etimologías más o menos rudimentarias que parecen ir ensanchando la brecha entre las palabras y las cosas.

Que buena parte de la investigación acerca de la naturaleza de lo mental se haya encaminado, de la mano de Wittgenstein, Ryle o Chisholm (*supra*), hacia el análisis del lenguaje psicológico no deja de revelarse como una grácil voluta de la historia si tenemos en cuenta que nuestra comprensión de los lazos entre los pensamientos y las cosas y nuestra comprensión de los lazos entre las palabras y las cosas parecen haber sido siempre la urdimbre y la trama de un mismo tejido. Es más:

---

[38] En ese contexto, como es fácil apreciar, hilvana Platón la concepción del lenguaje bajo la metáfora de la herramienta –el taladro y la lanzadera se mencionan expresamente– que tantos frutos guardaba para Wittgenstein.

[39] La idea de que no puede haber en realidad pensamientos falsos provendría, como se ha adelantado *supra*, de Parménides (*cf*. *Crátilo* 385*b*), tal vez por mediación de Antístenes, pero también de la concepción mágica de los nombres (Calvo 1983: 355, *cf*. Laín Entralgo (1897: 33, *supra*) y Ogden y Richards (1923: 52, *supra*); Sócrates mostrará en el diálogo que se deriva igualmente del convencionalismo de Hermógenes, inspirado en Protágoras, y del naturalismo de Crátilo, de raigambre heraclítea –hay a lo largo del diálogo constantes referencias a Heráclito, blanco pero también inspirador de la polémica.

como se ha visto, la cuestión del error –la perplejidad ante el hecho de que podamos tener pensamientos falsos, o errar al referirnos a algo, la constatación de que el error es un *explanandum* crucial si hemos de entender cuáles son los vínculos que unen a creencias o deseos con las cosas a las que atañen– ha acompañado también desde muy temprano a la reflexión sobre el significado lingüístico. Pero la reprobación final del lenguaje como ruta solitaria hacia el conocimiento (Calvo 1983: 358) es el amargo fruto del *Crátilo*: "si uno busca las cosas dejándose guiar por los nombres –examinando qué es lo que significa cada uno–, ¿no comprendes que no es pequeño el riesgo de dejarse engañar?" (*Crátilo* 436*b*); pues "[…] hay que conocer y buscar los seres en sí mismos más que a partir de los nombres" (*Crátilo* 439*b*). No es difícil ver aquí una advertencia –de aire, por lo remoto, casi oracular– contra ciertas lecturas del giro lingüístico de la filosofía reciente.

Lo que ahora nos ocupa, sin embargo, es la cristalización de la conciencia de que la capacidad del lenguaje, o del pensamiento, de designar o describir errónamente lo real es tan asombrosa como su capacidad de hacerlo límpidamente, sin extravío alguno. Un paso más nos invita a dar Giorgio Colli en su sosegada lectura de la ebullescente realidad intelectual griega de los siglos VI y V a.C.: el descubrimiento decisivo de los presocráticos sería, a su juicio, el de que "[…] el mundo de la apariencia es mera representación, conjunto de relaciones, y puede ser dominado tan solo por el conocimiento. Pero el mundo es ambiguo, mudable por su misma naturaleza representativa que multiplica indefinidamente las relaciones bajo las más variadas perspectivas, y ambiguos son necesariamente los conceptos que lo expresan" (Colli 1988: 35); "[…s]e descubre", así pues, "el más profundo dolor: toda expresión es inadecuada" (Colli 1988: 31). Lo sabemos, pero sabemos también –como Wittgenstein–, aunque no alcancemos a entenderlo del todo, que:

> La concordancia, la armonía entre pensamiento y realidad consiste en que cuando digo falsamente que algo es *rojo*, a pesar de ello no es *rojo*. Y [en] que cuando le quiero explicar a alguien la palabra 'rojo' en la proposición "Esto no es rojo", señalo a este fin algo rojo. (Wittgenstein 1953: §429)

Es, en suma, esa misma grieta que entre el pensamiento y las propias cosas –o entre ellas y nuestras palabras– hiende el decidido naturalismo jonio la que todavía torpemente tratamos de salvar. Así visto, no parece fortuito que el error –nuestros propios errores– se perfile como uno de los territorios en los que la cuestión se plantea con mayor nitidez, y en los que aparecen más descarnadas las dificultades que arrostran los numerosos intentos de darle respuesta: allí donde la distancia entre mente y mundo se acrecienta, confiamos en que acaso fugazmente.

RAÍCES Y DESARROLLO DE LA CONCEPCIÓN COGNITIVISTA DE LO MENTAL

Crisis y vigencia del conductismo

Ha acabado por convertirse en un lugar común, tanto en el relato historiográfico como en la conversación de expertos y legos, que la hegemonía del conductismo como concepción de lo mental –concepción, convendría matizar, que, en al menos algunos de los muchos rostros de la teoría, opera solo por vía negativa– y como concepción de la investigación psicológica se habría derrumbado con estrépito hace ya varias décadas, y que los principios que animaban el trabajo de los conductistas habrían quedado profunda y completamente desacreditados. El auge del conductismo aparece entonces como un mal menor, una enfermedad juvenil –"excesos de la juventud", apuntaba ya Angell (1913: 270), "necesaria adolescencia", reitera Mora (1992b: 108)– de la que aprendimos poco más que algunas prácticas higiénicas que aún observamos a la hora de teorizar –"labor de limpieza doméstica", decía Hebb (1968: 4)–, cuando no, en miradas menos benévolas, se perfila como un extraño episodio de enajenación transitoria, apenas comprensible como una especie de *folie imposée* por unos pocos visionarios sobre una comunidad científica desprevenida, aletargada quizá por los excesos especulativos de las escuelas introspeccionistas. Hay, desde luego, lecturas retrospectivas mucho más mesuradas, como la que nos ofrece Pujadas (2002: 11) en las líneas que abren su trabajo: el descrédito del conductismo queda enmarcado en un ciclo de aire kuhniano que abarca también la posterior caída en desgracia de las posiciones epistemológicas que lo propiciaron, y al término del cual cabe cierta reparación "bajo una nueva luz" de algunas de aquellas ideas tan tajantemente rechazadas, tras haber permanecido años "[…] arrinconadas como errores que casi incomprensiblemente alguien osó defender en tiempos que, aunque recientes, parecen irreversiblemente alejados".

Es cierto, con todo, que la idea de la actividad científica que trasluce en el pensamiento de Watson –y que alienta su empeño por "pensar hasta su extremo la exigencia de la objetividad en psicología" (Politzer 1969: 207)– es difícilmente respaldable hoy por hoy. En realidad, había quedado trasnochada ya en su día: en 1930, cuando diera a la imprenta la edición revisada de *Behaviorism* (Watson 1924), Watson seguía condenando –como Comte (1830-1842)– la introducción en las teorías científicas de referencias a fenómenos no observables; mientras tanto, en el seno del Círculo de Viena, se encontraba ya prácticamente madura una concepción de la ciencia, el positivismo lógico, que, al amparo de un cuidadoso estudio de la estructura lógica de las teorías, abriría las puertas a una proliferación de entidades teóricas –estrictamente definidas, eso sí, mediante reglas de correspondencia que las ligaran a un lenguaje observacional previo e independiente. Tales entidades teóricas, por lo demás, ya abundaban en la jerga en que se venía desplegando la investigación

física desde el advenimiento de la teoría atómica, y habían sido objeto del influyente análisis operacionista de Bridgman (1927), cuyas huellas están tan presentes en el pensamiento de los positivistas lógicos como en el de los propios conductistas. De hecho, como recuerda Boring (1950: 676), los planteamientos de Bridgman, un destacado físico de Harvard, bien podían leerse como un respaldo a los que Meyer (1911, 1921) había auspiciado en el ámbito de la psicología. Hace bien en señalar Rodríguez (2001a), a este respecto, que:

> Por muy saludables que en cierto momento de su historia hayan sido para la psicología las restricciones del conductismo metodológico, y por mucho que tengamos que celebrar la limpieza conductista de las pseudoexplicaciones psicológicas del sentido común, hay que decir que la concepción de la ciencia que formaba el núcleo de este movimiento no es ya la de nuestro presente […]. Son legítimas las entidades inferidas, cuando los procesos de inferencia se han controlado cuidadosamente, y cuando toleran predicciones independientes de las que las generaron. (Rodríguez 2001a: 95)

Ahora bien, la anotación –entrelazada con la anterior– de que "no todos los términos teóricos de una ciencia han de estar conectados directamente con los observacionales" (Rodríguez 2001a: 95) podría dar a entender que es la desacreditación del proyecto fisicalista del Círculo de Viena lo que debilita los cimientos del conductismo metodológico –aunque fuera sólo porque la jerga del vocabulario teórico y el observacional evoca vívidamente los análisis de los positivistas lógicos. Pero, sin detrimento de la verdad que esa afirmación pueda revestir en cuanto hace a los hechos históricos –pues no es aventurado atisbar en los devastadores ataques al positivismo lógico de Hanson (1958) o Kuhn (1962) el mismo *Zeitgeist* que impulsara a los pioneros del cognitivismo–, parece claro que la solidez epistemológica del conductismo de Watson estaba comprometida de antemano, incluso bajo los propios presupuestos de los positivistas vieneses. El *tópos* de la crisis del conductismo enlaza ésta con la caída en desgracia del positivismo lógico porque de los argumentos de –por ejemplo– Hempel (1935) se desprendía desde luego cierta afinidad con una concepción conductista de lo mental. Sin embargo, la idea medular del conductismo psicológico en Watson como en Skinner –el rechazo de la teorización sobre fenómenos no observables– tampoco convivía bien con el positivismo lógico. De hecho, ni siquiera un positivista como Hempel (1935), entonces un pensador de irreprochable rectitud reduccionista, vio prudente plegarse a las restricciones metodológicas dictadas por Watson. En efecto, supeditar la incorporación de la psicología al aparato fisicalista al fruto del método de Watson era, a ojos de Hempel, un proceder poco sensato, dado que:

> [...o]ne cannot expect the question as to the scientific status of psychology to be settled by empirical research in psychology itself. To achieve this is rather an undertaking in epistemology. (Hempel 1935: 16)

En cualquier caso, razón de más: la idea de la actividad científica que trasluce en el pensamiento de Watson es, efectivamente, difícilmente respaldable a día de hoy.

Aunque no exactamente por los mismos motivos, un diagnóstico semejante resulta acertado en cuanto a la vinculación entre el positivismo lógico y otras vertientes del conductismo. Como han recordado O'Donohue y Kitchener (1999: 8), el minucioso estudio de Smith (1986) resulta convincente en cuanto a que ni en el caso de Skinner, ni, con matices, en el de Tolman o Hull, es verdad que la epistemología de fondo viniera importada tal cual del Círculo de Viena, ni que el desarrollo de las teorías conductistas estuviera subyugado a esa epistemología prestada, ni, en suma, que la supuesta tosquedad de dicha epistemología implicara un quebranto irreparable para las construcciones teóricas de los conductistas o para sus proyectos de investigación. Antes bien, como señala Ringen (1990) después de detallar el compromiso del Círculo de Viena con la naturaleza lógica de la teoría de la ciencia, al que apela Hempel (1935, *supra*) en su rechazo del proyecto de Watson:

> Radical behaviorism is orthogonal to these positivist doctrines. It should, then, come as no surprise that attempts to portray radical behaviorism as a form of logical positivism occasion some fundamental misunderstandings. (Ringen 1990: 165)

En particular –si Ringen (1990) está en lo cierto– las razones del antimentalismo incansablemente predicado por Skinner habrían quedado una y otra vez ofuscadas por la asunción acrítica de ese parentesco entre su posición y la de los positivistas lógicos. Las convicciones antimentalistas de Skinner no descansan sobre una concepción de la estructura lógica y la justificación de las teorías científicas –como, por ejemplo, en Hempel (1935)–, sino más bien sobre una visión de la labor del científico estrictamente ceñida al *Novum Organum* del canciller Francis Bacon (1620) –empirista, descriptivista, inductivista, pragmática y profundamente antiteórica. Más exactamente –de acuerdo con Ringen (1990: 167)– el antimentalismo de Skinner hiende como una cuña sus certidumbres baconianas: reposa sobre su empeño en la predicción y el control de la conducta –la visión pragmática de la ciencia que compartía con Watson– a la vez que sirve de base a su escepticismo ante la teorización. Así pues, a Skinner no le habrían repugnado las explicaciones mentalistas debido a su talante antiteórico; por el contrario, dicho talante habría germinado en él debido a su desconfianza en la utilidad práctica de las explicaciones mentalistas. En cualquier caso –como también registra Ringen (1990: 164)–, el propio Skinner (1945: 277) había rechazado abiertamente el positivismo lógico, encarnado en el afán de desarrollar una lógica de la investigación científica que, a juicio de Skinner, no podía sino quedar subsumida en la venidera ciencia empírica de la conducta de los propios científicos. Pero incluso de no haber sido así, habría que dar la razón a Savin (1980) cuando a fin de recalcar la mutua independencia del positivismo lógico y el conductismo psicológico, sea de índole metodológica o "radical" apuntaba que:

> […] Hempel's thesis about the ultimate foundations of psychology has no obvious bearing on the day-to-day activity of psychologists, while Skinner's advice about how psychologists can best spend their time has no obvious epistemological content. (Savin 1980:11)

Más atinado parece –como hace Yela (1980/1996: 171)– ver en la búsqueda de "[…] unas reglas explícitas y formales para elaborar conceptos, enunciados y teorías […]" un vínculo entre el conductismo sistemático, o neoconductismo, de Tolman o Hull y el positivismo lógico, o neopositivismo, de Carnap o Hempel. Vale, con todo, la tesis de Smith (1986): los neoconductistas no partieron de tomar al pie de la letra los preceptos del los neopositivistas para después aplicarlos en sus laboratorios. En ocasiones, de hecho, es a los propios conductistas a quienes se adeudan determinadas nociones epistemógicas, como la de variable intermedia, o interviniente, que había acuñado Tolman (1932, 1936) advirtiendo de que el significado de cada una de esas variables residía en la expresión de las regularidades entre las variables observables que modulara.

Aun así, la conclusión de O'Donohue y Kitchener (1999: 8) –"[…i]f this is correct, then the demise of logical positivism did not carry with it the demise of behaviorism"– es a todas luces precipitada, por cuanto atañe a una cuestión fáctica de la historia del pensamiento psicológico pero parte de premisas que, como mucho, serían firmes respecto de las relaciones lógicas entre diversas estructuras conceptuales. Dicho de otro modo: que la repulsa del positivismo lógico no implicara la del conductismo puede ser tan verdad como que, de hecho, una cosa llevó a la otra.

Es cierto también, por otro lado, que los frutos de décadas de investigación psicológica imbuida de nociones y intereses cognitivistas son copiosos –*pace* Skinner (1987)–, y que constituyen una prueba palpable de que al menos algunas de las restricciones preconizadas por el conductismo eran desmedidas. Incluso el experimentalista más reacio a la reflexión filosófica puede ser sensible a esta circunstancia. Acaso la actitud de muchos psicólogos ante la figura antes tutelar de Skinner quede bien reflejada en las palabras de Stich (1998: 649), quien parece casi querer excusarse de una infidelidad cuando apunta que:

> […] we now have an enormous collection of experimental data which, it would seem, simply cannot be made sense of unless we postulate something like "information processing mechanisms in the heads of organisms" […].

El conductismo –se diría– pareció en su momento una buena idea –quizá el cognitivismo, incluso, pareciera una aventura un tanto insensata–, pero han ido pasando los años y aquí están todos estos resultados que sería imperdonable ignorar. La prosperidad del proyecto cognitivista de investigación –en suma– haría cada vez más indefendibles las severas cortapisas metodológicas que el conductismo pretendiera imponer. Pero aun concedidas estas certezas, no menos cierto parece que la concepción cognitivista de la mente y de la investigación psicológica arraiga tan hondo en el conductismo que resulta inexcusable, al tratar de exhumar esas raíces, poner de relieve cuanto había de clarividente en los incendiarios planteamientos de los conductistas, tan desacordes como pudieran resultar con nuestras intuiciones

espontáneas –pues cuanto había en ellos de obcecado es más patente hoy. Ahora bien: aun si se da por bueno este intento de reparación –"bajo una nueva luz", como escribe Pujadas (2002: 11, *supra*)–, conviene no confundirlo con uno de reescritura histórica. En definitiva, como se ha apuntado ya, nada en estos razonamientos estorba a la conclusión de que el conductismo –desde la óptica de la sociología o la historia de la ciencia– presentara tal o cual perfil, o de que su defenestración en favor del cognitivismo tuviera lugar bajo el signo de tales o cuales fuerzas.

Algunas voces han cuestionado incluso, acaso con más esfuerzo que repercusión, que el auge del cognitivismo y la defunción del conductismo hayan de hecho ido parejas. No podrían mostrarse más tajantes en este sentido O'Donohue y Kitchener (1999: 7): "It is a myth that the cognitive revolution killed behaviorism". La minuciosidad en la descripción de los múltiples nichos académicos y aplicados en que el conductismo sobrevive, con todo, no viene acompañada en el argumento de igual exactitud en la exégesis del mito que se pretende desarticular. Pocos estudiosos –incluso pocos apologetas– del cognitivismo *defenderían* la tesis de que la revolución cognitiva acabó con el conductismo en el sentido de que lo haya expulsado totalmente de la psicología –ni de sus estructuras conceptuales ni de las sociales–, aunque puedan *afirmar* tal cosa con propósitos retóricos. Es saludable, por supuesto, reclamar comedimiento en esos afanes retóricos, y argumentos como los de O'Donohue y Kitchener (1999) ayudan a hacerlo. Ellos mismos –por otro lado– han hecho ver con notable ecuanimidad que el intenso tinte emocional de ciertas reacciones contra el conductismo pudo en buena medida deberse a la "superlativa" retórica de Watson, Skinner o Ryle (O'Donohue y Kitchener 1999: 9). Pero también es saludable reparar en que, una vez desarmada cierta retórica, conviene no dejarse arrastrar por la de signo opuesto: que la transición de la hegemonía conductista a la cognitivista no haya consistido en el exterminio de todo rastro del conductismo no implica que no se haya producido una transición. De hecho, el mito que O'Donohue y Kitchener denuncian no es sino uno de sus síntomas más patentes.

Precisamente en el contexto de un estudio de la transición de la supremacía del conductismo a la del cognitivismo, Estany (1999) –contra los argumentos de Zuriff (1985) o de O'Donohue y Kitchener (1999: 2, *infra*), contra, también, los de MacKenzie (1977), de ánimo más adverso hacia los planteamientos conductistas– da por buena una cierta homogeneidad entre las escuelas conductistas, así como el carácter anómalo de determinados resultados empíricos y trabajos conceptuales:

> A pesar de las diferencias entre Watson y los neoconductistas (Hull, Tolman, Skinner, entre otras), los trabajos de todos estos psicólogos constituyen un nuevo período de ciencia normal hasta las primeras anomalías, que Palermo [1971] sitúa en una tesis de Kuenne (discípul[a] de Spence) sobre la conducta de transposición aplicada a los niños. A Kuenne [1946] le siguen Kendler [...y] Kendler [1962], Postman [1963], Harlow [1949, 1953, 1958] y Lashley [1951] en el cuestionamiento de algunos de los supuestos centrales del conductismo. Especial mención merece Chomsky [1959] con su respuesta a *Verbal Behavior* de Skinner [1957]. (Estany 1999: 138)

El caso de Lashley (1951), por lo demás, venía revestido de una formidable potencia simbólica: Lashley había sido discípulo de Watson en la Universidad Johns Hopkins, en Baltimore, Maryland, y uno de sus más estrechos colaboradores durante años. Uno de los detonantes de la deserción de Lashley, además, atañía a una clave de arco del modelo teórico construido por Watson –su concepción de las relaciones entre el pensamiento y el habla. En efecto, Lashley se mostraba sumamente incrédulo ante la tesis de que cada palabra emitida por el hablante se convierte en el estímulo que elicita la siguiente. Ya el propio Watson (1924, 1930) se había visto obligado a retractarse de su osadísimo planteamiento inicial en este punto –a saber: el pensamiento es un conjunto de hábitos laríngeos– al reparar en que incluso una laringectomía total dejaba intacta la capacidad cognitiva de los pacientes. Así pues, la disidencia del viejo alumno y colega tocaba sobre llaga. La ponencia de Lashley –la "más iconoclasta y memorable", según el juicio de Gardner (1985: 26), de las que se pronunciaran unos años antes –en septiembre de 1948– en el marco del simposio sobre los mecanismos cerebrales de la conducta que albergaba el Instituto de Tecnología de California bajo el auspicio de la Fundación Hixon–, reintroducía las ideas de control central de la conducta y de planificación jerarquizada. En torno al trabajo de Lashley "[…] cristalizó una creciente conciencia, por parte de muchos científicos sensatos, de que la adhesión a los cánones conductistas estaba volviendo imposible el estudio científico de la mente" (Gardner 1985: 28)[40].

[40] Aunque la distancia histórica era ya demasiada como para que eso pesara en los sentimientos hacia el conductismo, no deja de resultar interesante que el problema del orden secuencial de la conducta, que ahora suscitaba con nitidez la necesidad de apelar, contra la ortodoxia conductista, a sistemas de planificación y control central, hubiera sido empleado casi doscientos años atrás, en una obra en la que muchos creyeron ver una cierta prefiguración de las tesis conductistas, contra la atribución al alma del gobierno voluntario del cuerpo. En efecto, su encendida disputa contra esa idea, defendida entonces por Georg E. Stahl (1715) con el mismo ardor con que defendía que la noción de *flogisto* podía dar cuenta de fenómenos tan dispares como la combustión y el color, llevó a Julian Offray de la Mettrie a apelar, en un párrafo de *El hombre máquina* que vale la pena citar *in extenso*, al ejemplo de un músico:

> Basta con arrojar la mirada sobre un violinista. ¡Qué ligereza y qué agilidad en los dedos! Los movimientos son tan rápidos que casi parece no haber sucesión. Luego, ruego, o más bien desafío a los staahlianos [*sic*] a que me digan, ellos que conocen tan bien todo lo que nuestra alma puede, cómo sería posible que ésta ejecutase tan de prisa tantos movimientos, y movimientos que tienen lugar tan lejos de ella y en lugares tan diversos. Esto sería como imaginar a un flautista que pudiera ejecutar brillantes cadencias sobre una infinidad de agujeros que no conociera, y a los que ni siquiera pudiera aplicar los dedos. (de la Mettrie 1748: 38-39)

La controversia entre las tesis de Albrecht von Haller –quien, pese a sus muchos desencuentros con él, seguía de cerca en este aspecto a Stahl– y las de Robert Whytt, quien se mostraba convencido de que un alma responsable del pensamiento y de la vida –al modo aristotélico– se extendía por todo el cuerpo adoptando diversas funciones según qué órgano la albergara, seguía palpitando en 1870, cuando Huxley, en su charla acerca del alma y la vivisección de sapos, toma partido por la posición de Whytt –abanderada entonces por un alumno de du Bois-Reymond, Eduard Pflüger– aunque sólo para abocarla al esperado desenlace cartesiano: "[…] I am unable to see in what respect the soul of the frog differs from matter" (Huxley 1870: 7). Ya entonces, curiosamente, la posición de von Haller o Stahl queda descrita, sin nombrarlos, como aquella en la que se conciben las operaciones del alma sobre el

Con un sinsabor parecido –pero de ecos menores– acabaría topándose Skinner, andando el tiempo. Precisamente en 1951, al tiempo que el trabajo de Lashley llegaba a la imprenta, Keller y Marian Breland, que colaborarían largo tiempo con Skinner como amaestradores de palomas, hacían gala de su optimismo respecto a la posibilidad de aplicar los principios conductistas en sus quehaceres fuera del laboratorio. Pero al cabo de diez años, los propios Breland y Breland (1961) cuestionaban abiertamente la irrelevancia de las variables internas. Como anota Mora (1992b: 85), "[…l]a sinceridad de los colaboradores rompía los dogmas del maestro".

Aunque sus repercusiones se harían notar más lentamente, fue también en 1951 cuando Kurt Lewin publicó una tajante refutación intuitiva del intento de construir una explicación completa de la conducta humana a partir de mecanismos de condicionamiento operante. Con el ánimo de convocar en su favor la fuerza de la argumentación de Lewin, George A. Miller, Eugene Galanter y Karl H. Pribram (1960) recordarían como en el caso de una persona que trata de echar una carta al correo –el ejemplo elegido por Lewin–, y de acuerdo con la doctrina del condicionamiento operante,

> […] el efecto de haber echado la carta al buzón tendría que haber sido reforzar la asociación entre los buzones, por un lado, y la respuesta de sacar del bolsillo la carta que hay que echar, por otro. Nuestro pobre amigo tendría que haber llevado a cabo respuestas no consumadas ante tres o cuatro buzones antes de que hubiera disminuido la fuerza de la asociación. Sin embargo, en vez de acumular la fuerza del hábito, de hecho ya no mostró más interés por los buzones. (Miller, Galanter y Pribram 1960: 70)

Como sucedería con frecuencia, las refriegas entre conductistas y gestaltistas servían de acicate al desarrollo del cognitivismo, que no sería desacertado entender, en parte, como una asimilación del pensamiento gestaltista en el seno de criterios metodológicos heredados del conductismo. Eso, desde luego, es lo que parecen estar haciendo Miller, Galanter y Pribram al reorientar las conclusiones que Lewin extraía del ejemplo hacia su propio esquema de planes e imágenes, apartándolas de la idea de cuasi-necesidad que Lewin (1951) había dejado bosquejada en el contexto de su noción de espacio vital.

En el seno de la psicología social, donde la influencia de Lewin sería más duradera, la teorización basada en la postulación de estados internos capaces de dirigir la conducta constituía, en palabras de Leahey (2005: 388), "[…] una vigorosa

---

cuerpo como las de un músico "sobre un órgano o algún otro instrumento" (Huxley 1870: 4), quizá en la estela de la metáfora, abiertamente rechazada por Descartes (1641/1642: VI), del alma como el timonel del buque corporal.

En su conferencia de 1951, Lashey elegiría ilustrar su posición con el caso de un pianista. En la confluencia de la metáfora de von Haller o Stahl con el ejemplo esbozado por de la Mettrie y luego elaborado por Lashley, así pues, el hombre y el homúnculo se encuentran en un juego de espejos: la delicadeza con la que el músico interior, el homúnculo, toca su instrumento, el cuerpo, se destila en aquella con la que el músico exterior, el hombre, arranque del suyo la armonía buscada. Tanto Lashley como de la Mettrie, desde luego, trabajaban por la desahucio del homúnculo.

psicología cognitiva fuera de la órbita del conductismo estricto". Particularmente pujantes resultaron las investigaciones de Festinger (1957) y Festinger y Carlsmith (1959) sobre lo que dio en llamarse "disonancia cognitiva" –tensiones internas al sistema de creencias y actitudes de un sujeto que provocan una tendencia a ajustar dichas creencias y actitudes de suerte que se reduzca la tensión. Pese a que los detalles de la interpretación de Leahey son susceptibles de objeciones sustanciales[41], es difícilmente cuestionable que el esqueleto conceptual de la teoría de la disonancia cognitiva está mucho más cerca del incipiente cognitivismo que, siquiera, de un conductismo mediacional. Las severas directrices del programa de investigación delineado por Watson, al igual que las elaboradas construcciones teóricas de Tolman o Hull, contrastan vivamente con la desenvoltura con que Festinger introduce en sus experimentos y explicaciones referencias a creencias, actitudes o valoraciones afectivas.

A la luz de todo esto, no es inusual que la crisis del conductismo quede retratada –en esa historiografía heredada– como el estrepitoso desmoronamiento de un proyecto de investigación asfixiado por insostenibles restricciones metodológicas, al mismo tiempo que se reconoce la fundamental continuidad de los planteamientos metodológicos del cognitivismo respecto de los adelantados por los conductistas. Pero si el cognitivismo adoptó, al menos en lo esencial, la metodología que había prosperado al amparo del conductismo, entonces resulta sensato pensar que los motivos de la crisis del conductismo y del auge del cognitivismo no fueron, en lo esencial, de índole metodológica. Parece dar razón a ese argumento la investigación emprendida por Mora (1992b) en torno a lo que él considera "contradicciones

[41] Así, que Festinger entendiera las creencias "[…] en términos del sentido común" es cierto sólo en el sentido de que éstas "[…] controlan la conducta" (Leahey 2005: 388), pero la naturaleza inconsciente tanto de las tensiones internas que constituyen la disonancia como de los esfuerzos por aminorarla sólo ha podido formar parte del sentido común, obviamente, después de que las investigaciones de Festinger nos la hicieran ver. Nada de esto, sin embargo, pone trabas a la conclusión de que la teoría de Festinger sea propiamente cognitiva, puesto que nada compromete al cognitivismo a aceptar únicamente estados o procesos psicológicos conscientes como términos teóricos (*cf.* Malcom 1971, *infra*) –ni siquiera si aceptamos la lectura del cognitivismo como reivindicación de la psicología del sentido común que propone, *inter alia*, Fodor (1968), la cual sólo implica que creencias y deseos conscientes tienen cabida en el vocabulario teórico de la psicología, no que determinados estados psicológicos inconscientes no puedan tenerla.

Por otra parte, que los fenómenos experimentales descritos por Festinger y Carlsmith contravinieran la ley del efecto, como afirma Leahey (2005: 388), sólo puede darse por verdadero, y aún así asumiendo ciertas distorsiones, si por ello se entiende lo que Thorndike denominaba 'efecto extendido', donde son conductas temporalmente adyacentes a la que ha producido consecuencias satisfactorias las que ven incrementada su probabilidad de ocurrencia –*nota bene*: su probabilidad de ocurrencia, no su valoración afectiva como en los trabajos de Festinger y Carslmith. En todo caso, el cuestionamiento de la ley del efecto habría sido al fin y al cabo una noticia grata para conductistas como Watson, que siempre la rechazó –*cf.*, por ejemplo, Watson (1914)– precisamente por encontrarla veladamente mentalista, y que sin duda volvería a ver en esa contradicción la misma indigencia de la psicología subjetiva que encontrara en su día en la polémica sobre el pensamiento sin imágenes que enfrentó a Külpe con Wundt y Titchener. Thorndike, dicho sea de paso, dio contestación a las críticas de Watson en un brillante artículo en *The American Journal of Psychology* (Thorndike 1927).

internas del conductismo skinneriano" –en realidad, resultados empíricos que cuestionaban determinados supuestos teóricos, anomalías kuhnianas como las espigadas por Estany (1999, *supra*). A la luz de su minucioso recuento, Mora (1992b: 78) concluye que "[…] la crisis del conductismo y del neoconductismo ha sido fundamentalmente una crisis de sus supuestos teóricos, no tanto de los fenómenos investigados o de su forma de investigarlos". Exacto: es razonable que a una crisis de supuestos teóricos siga una reformulación de tales supuestos –o incluso de algunos de orden preteórico, *cf.* Estany (1999: 86, *infra)*–; sería difícil encontrarle sentido, en cambio, a un proceso histórico en el que una crisis de directrices metodológicas dejara prácticamente intactas dichas directrices, pero diera lugar a una reformulación de supuestos teóricos. En esto al menos, parece que acaso sí cupiera esperar cierta concomitancia entre la lógica de la investigación científica y su despliegue histórico –a falta de indicios de alguna otra fuerza que pueda haberlas hecho discordar.

Más aquilatado que la historiografía heredada –aunque tal vez, como toda partición del tiempo histórico, también algo artificioso– parece el relato de la crisis del conductismo que esquematizara Yela (1980/1996), donde se diferencian una fase de crisis, una de declive y una de caída. La crisis, en la reconstrucción de Yela, habría venido originada por el cuestionamiento tanto de la validez de los canónes metodológicos establecidos –en Hebb (1960), por ejemplo– como de su cumplimiento efectivo –en Estes *et al*., eds. (1954). Durante la fase de declive, la conducta perduraría como objeto de estudio, pero irían quedando en tela de juicio algunos principios de interpretación de los resultados experimentales –así, a partir de Egger y Miller (1962), Kamin (1969) y Wagner (1969), por ejemplo, el condicionamiento pavloviano tenderá a interpretarse como signo de una reducción de la incertidumbre acerca de qué tipo de evento sucederá al estímulo condicionado: de la incertidumbre, claro está, *para* el organismo, o desde *su perspectiva*. Al cabo –dicho sea de paso– la reinterpretación del condicionamiento habría de convertirse en uno de las simientes de la psicología cognitiva: aún en un trabajo de sistematización metateórica relativamente tardío, como es Pylyshyn (1984), desempeña un papel primordial en el hilo de la argumentación la idea, tomada de Brewer (1974), de que:

> […] an account of human conditioning experiments that makes use of the notion that a subject is being informed of what will happen, or what he or she is expected to do, provides a better explanation of the observed phenomena than does an account based on reinforcement contingencies. (Pylyshyn 1984: 6)

O, con mayor énfasis si cabe en el carácter cognitivo de la explicación propuesta, que Pylyshyn pone de relieve con una insistencia que sin duda habría exasperado a Skinner:

> […] the most plausible explanation of human-conditioning phenomena is one given in terms of change in belief. It is the most straightforward explanation of what reinforcers do: they inform the subjects of the contingencies so subjects can select a course of action based on their beliefs and utilities. (Pylyshyn 1984: 216)

Finalmente, la caída del conductismo –en la que la conducta queda degradada del rango de objeto de estudio único, si bien se le confiere el de medio crucial de contraste de hipótesis– coincidiría con los albores del cognitivismo. En las postrimerías del conductismo se apagaba –cree Yela– "el intento más ambicioso y tenaz de toda la historia de la psicología […] de construir un sistema científico estrictamente lógico y objetivo […]" (Yela 1980/1996: 165).

Particularmente tajante es, respecto a la continuidad entre el enfoque de los conductistas y el de los cognitivistas, el dictamen de O'Donohue y Kitchener (1999: *xx*): "In short, it is our view that no one has yet set out the features distinguishing behaviorism from cognitivism". En realidad, ni siquiera los rasgos distintivos del conductismo como tal habrían quedado hasta la fecha convincentemente elucidados según el diagnóstico de O'Donohue y Kitchener (1999: 2). Su apuesta, inspirada en la tesis de Zuriff (1985), es que, dado que no parece existir un inventario de propiedades necesarias y suficientes que delimite el conjunto de las posiciones conductistas, convendría tratar de elaborar el propio concepto de *conductismo* bajo el modelo de la noción wittgensteiniana de aire de familia (Wittgenstein 1953), o de la teoría de propotipos de Rosch (1978). En todo caso, la constatación de la heterogeneidad del conductismo no es nueva, aunque una historiografía afín al cognitivismo pueda haberla emborronado. Ya en la segunda edición de la *Historia de la psicología experimental*, que publicara inicialmente en 1929, Boring había anotado, cuando se disponía a describir los días de hegemonía del conductismo en la psicología académica estadounidense –días que él circunscribía a la década de 1920– que por aquel entonces "[…t]odos eran conductistas […] y ningún conductista estaba de acuerdo con otro" (Boring 1950: 667). El propio carácter paradigmático del conductismo ha sido cuestionado por MacKenzie (1977), quien sospecha que tras las divergencias entre los neoconductistas se esconde una constelación de escuelas preparadigmáticas y no un paradigma maduro. Más recientemente, Gondra (1992) –apoyándose en el concienzudo recuento de Quintana (1985)– recordaba cómo:

> […] los "primeros conductistas" formaban un grupo heterogéneo con intereses y procedencias muy diversas, a los que sólo les unía la fe en el objetivismo científico y la oposición al método de la introspección. […] Los neoconductistas compartían la misma definición de la psicología como ciencia objetiva de la conducta, pero en lo demás mantenían serias divergencias. (Gondra 1992: 15-16)

Igual de taxativo se muestra Yela (1980/1996: 172), que anota que más allá de la homogeneidad metodológica "todo lo demás se quiebra y se fracciona" –de "[…] la torre de Babel conductista" nos habla sin ambages Leahey (2005: 363). Como concisa y convincentemente detalla el propio Yela:

> Ni Watson ni los conductistas, ni estos entre sí, comparten un cuerpo común de conocimientos, explicaciones y resultados fundamentales, que pudiera ir progresando y se articulara, por fin, como se pretendía, en *una* psicología conductista.

> […] Discrepan en cuanto a lo que el animal aprende: respuestas, conexiones estímulo-respuesta (S-R), asociaciones entre estímulos (S-S), expectativas, relaciones. Discrepan en cuanto al mecanismo por el que el animal aprende: contingüidad, reforzamiento, ensayo y error vicario, confirmación de expectativas, transposición. Y discrepan en cuanto a la interpretación de ese mecanismo: muestreo de estímulos y respuestas en el establecimiento incremental de conexiones entre los patrones de energías y de movimientos, o refuerzo como reducción de necesidades, reducción de impulsos, satisfacción hedónica, mantenimiento de la propia actividad, cambio significativo en la estructura de la estimulación o mera comprobación empírica del aumento de la probabilidad de la respuesta. (Yela 1980/1996: 172-173)

Que el advenimiento del cognitivismo deba entenderse como una revolución científica –en el sentido prefigurado por Brinton (1938) en su análisis del concepto de revolución política bajo la metáfora del proceso febril, y luego fijado por Kuhn (1962)– ha sido objeto de controversias tan pertinaces como madrugadoras. Contra el análisis kuhniano de la caída del conductismo elaborado por Palermo (1971), la propia Estany (1999) recuerda –por ejemplo– las críticas de Warren (1971): la supuesta revolución cognitiva sólo habría tenido lugar en los Estados Unidos, ya que el pensamiento psicológico europeo estaba dominado por la Gestalt, Piaget, Luria y Vygostki. Así las cosas, más que de una revolución cognitiva convendría hablar, como hace Rivière (1991b: 130), de la "anomalía histórica" del conductismo, o de cómo "[…] la psicología volvió a ser cognitiva" (Rivière 1991b: 132)[42]. En la misma línea incide incide Mandler (2002), quien además apunta la deliciosa anécdota relatada por George Miller y recogida por Baars (1986: 212): tras una conferencia contra el conductismo que Miller dictó en Oxford en 1963, alguien le hizo notar que sólo había tres conductistas en Inglaterra, y que ninguno se hallaba presente en la sala. Los brillantes estudios de Bartlett, el primer profesor de psicología experimental de la Universidad de Cambridge, sobre el empleo de *esquemas mentales* en el recuerdo de textos narrativos (Bartlett 1932), cuya aproximación a la investigación psicológica encontró continuidad en los trabajos sobre atención selectiva y memoria a corto plazo de Broadbent (1958), así como la noción de modelos mentales a la que Kenneth J.W. Craik (1943) había dado forma, parecían ejercer una influencia más poderosa sobre la comunidad psicológica del Reino Unido que los argumentos de Ryle (1949)[43].

---

[42] Un llamativo ejemplo del fenómeno expuesto por Mandler y Rivière puede hallarse en Frijda (1967), un temprano intento de articular las relaciones entre programas y teorías en el ámbito de la simulación computacional de procesos psicológicos (*cf. infra*). A modo de ejemplo de la aproximación a la simulación cognitiva que defiende –en la que no basta con la semejanza entre la conducta observada de los sujetos y del modelo computacional, sino que se busca también la verosimilitud en cuanto atañe a los procesos internos– Frijda, profesor de psicología en la Universidad de Amsterdamm, describe una investigación en curso sobre la capacidad de ofrecer definiciones verbales de conceptos en la cual la introspección sistemática, "a la manera tradicional de Würzburgo" (Frijda 1967: 63), desempeña un papel protagonista. Oswald Külpe había fallecido en 1915.

[43] Hasta veinte años más tarde no son palpables las repercusiones del trabajo de Craik, a quien citan expresamente Newell y Simon (1963), en las líneas de investigación que acabarían por confluir en el auge del cognitivismo. Otros veinte años después sería Johnson-Laird (1983, *infra*) quien reivindicaría la figura de Craik.

A su vez, las objeciones de Warren habían encontrado su réplica en una tajante puntualización de Weimer y Palermo (1973): el modelo de Kuhn relativiza el concepto de paradigma a su aceptación por parte de una determinada comunidad científica, por lo que la crítica es errada –al menos en la medida en que la comunidad psicológica europea continental y la anglosajona se mantuvieran aisladas durante la primera mitad del siglo. Por otra parte, Estany (1999: 196) impugna la presunta continuidad entre el cognitivismo y el estructuralismo wundtiano, contraponiendo una vaga –y cuestionable– coincidencia en el objeto de estudio –los fenómenos de la conciencia– a las diferencias metodológicas y de caracterización de éste; la tesis de que el cognitivismo guarda en realidad continuidad metodológica con el conductismo (*cf*. Thagard 1992, *infra*) se impone entonces de modo natural. En todo caso, mayor coincidencia cabría reconocer –argumenta Estany (1999)– entre el cognitivismo y la psicología de Titchener o de Külpe; es la misma puntualización que ya hiciera valer Baars (1986: 38).

Menos dudas alberga Haugeland (2002a) respecto a que el cognitivismo sea un paradigma kuhniano con todas las de la ley. Desde su punto de vista, que coincide en lo fundamental con la narración historiográfica hegemónica:

> Cognitive science –as contrasted with its predecessors, behavioral science and cybernetics– emerged in the mid-1950s, with the seminal works of Miller, Galanter, and Pribram (1960), Newell and Simon (1958), and Chomsky (1968), among others. […] In the seminal works just mentioned, the new science had a paradigm in the proper Kuhnian sense: a body of concrete scientific achievement that was "sufficiently unprecedented to attract an enduring group of adherents away from competing models of scientific activity" and "sufficiently open-ended to leave all sorts of problems for the redefined group of practitioners to resolve" (Kuhn 1962: 10). (Haugeland 2002a: 24)

En el contexto de una laxa aplicación del modelo de revolución política de Brinton (1938) al caso del cognitivismo, también Bialystok (1997) da por bueno que el paulatino abandono de la ortodoxia conductista en favor de las nuevas rutas teóricas anunciadas por el cognitivismo fue el resultado del choque entre dos paradigmas. De acuerdo con su análisis, además, las objeciones que contra la concepción cognitivista de lo mental se han ido acumulando desde entonces son connaturales al desarrollo de las revoluciones: los rebeldes de los primeros tiempos acaban por usurpar las esferas de poder que ellos mismos asaltaron, y el descontento se va enseñoreando de las generaciones que les suceden. Así que la hegemonía de la psicología computacional sería hoy por hoy –según el dictamen de Bialystok– el Palacio de Invierno[44].

[44] El carácter paradigmático del funcionalismo ha sido también discutido por Antonio Blanco Salgueiro (2001) y Pascual F. Martínez-Freire (2001). Tras apuntar que Lycan (1994) hace suya la tesis de Palermo (1971) y describe al funcionalismo como el "paradigma reinante en filosofía de la mente", Blanco (2001: 106) deja constancia de su impresión de que, de ser así, el auge del funcionalismo habría sido en todo caso algo más breve de lo que las palabras de Lycan darían a entender, con signos de declive ya desde principios de la década de 1980; Martínez-Freire (2001: 92, 113), por otra parte,

En cualquier caso, es un rasgo bien conocido de la concepción kuhniana de la historia y la sociología de la ciencia que no es la crisis de un paradigma –por severa que sea la acumulación de anomalías– lo que da pie a un periodo de ciencia extraordinaria, sino más bien la perspectiva de que un paradigma alternativo pueda hacerse fuerte en las debilidades del antiguo. Del mismo modo, según el argumento de Kuhn (1962: 150-151), que, dado que "[…l]as revoluciones políticas tienden a cambiar las instituciones políticas en modos que esas mismas instituciones prohíben", su capacidad de generar tales cambios "depende de que [las revoluciones] sean sucesos parcialmente extrapolíticos o extrainstitucionales", cuando una revolución científica desarbola un paradigma es de esperar que su origen se halle fuera de la órbita del propio paradigma. Al amparo precisamente de estas observaciones sostiene Campos-Roldán (1999) que el giro cognitivo es exterior al desarrollo del conductismo por mucho que las sucesivas reformulaciones del proyecto de Watson pudieran haberlos aproximados. Desde el mismo punto de vista, Estany (1999) jalona los inicios de la ciencia cognitiva –siguiendo de cerca a Gardner (1985)– en torno al *Hixon Symposium* de 1948, el *Symposium on Information Theory* de 1956, que reunió a Chomsky, Miller, Newell y Simon en el Instituto de Ingeniería Eléctrica y Electrónica, perteneciente al Instituto de Tecnología de Massachusetts (MIT) –ocasión a la que también aluden Baars (1986) y Mandler (2002). Poco antes del encuentro en Massachusetts, que tuvo lugar a principios de septiembre, Newell y Simon habían asistido también a la *Summer Research Conference on Artificial Intelligence*, un encuentro de trabajo que John McCarthy había organizado en Darmouth College y que se extendió durante la mayor parte del verano. Además, también en la estela de Gardner (1985), Estany (1999) hace destacar el papel dinamizador, en la década de los setenta, de la Alfred P. Sloan Foundation de Nueva York. Como textos referenciales, cita, como es canónico, los trabajos de Miller (1956), Bruner, Goodnow y Austin (1956) y Miller, Galanter y Pribram (1960), además de la reseña de *Conducta Verbal* (Skinner 1957) que Chomsky publicó en *Language* en 1959. Hasta comienzos de la década de 1970 habría que esperar, sin embargo, para ver acuñado como nombre de esa confluencia interdisciplinar el término "ciencia cognitiva": ya en 1973 lo emplea Hugh Cristopher Longuet-Higgins, del Departamento de Psicología Teórica de la Universidad de Edimburgo, en su réplica al informe sobre los progresos en el ámbito de la inteligencia artificial que James Lighthill había presentado al *Science Research Council* del Reino Unido –un informe

rechaza frontalmente la idea de que el funcionalismo constituya un paradigma kuhniano en el seno de la ciencia cognitiva, para defender en cambio que el funcionalismo es una hipótesis filosófica que permite "[…] consolidar la unificación de las ciencias cognitivas, ya unificadas por la hipótesis empírica del sistema de símbolos" físicos. Receloso de ligar la suerte de la ciencia cognitiva al funcionalismo –como Hempel (1935: 16, *supra*) de hacer depender la unidad de la ciencia del proyecto de Watson–, Blanco se muestra partidario de deslindar funcionalismo y cognitivismo y abrir así la puerta a un cognitivismo no funcionalista, acaso postfuncionalista o incluso antifuncionalista. Entre las dificultades del funcionalismo que se apuntan como motivos para esa desconfianza ocupan un papel destacado el problema del contenido mental y el problema de la eficacia causal de lo mental, en los que se centre la presente investigación.

muy crítico, que conllevó la retirada de buena parte de los fondos que las autoridades británicas venían destinando a la investigación en ese terreno (Russell y Norvig 2003: 22) –; poco después lo utilizarían también, como ha señalado Martínez-Freire (2001: 89), Daniel Bobrow y Allan Collins en *Representation and Understanding. Studies in Cognitive Science*, un trabajo colectivo editado por ellos en 1975 cuyo prefacio comenzaba anunciando que "Este libro contiene estudios en un nuevo campo que llamamos *ciencia cognitiva*".

Erraríamos, sin embargo, al obviar el hecho de que buena parte de los desarrollos que darían alas al cognitivismo llegaron de la mano de investigadores provenientes de las filas conductistas –cuando cruza uno la linde y abandona los predios del propio paradigma sería sin duda una cuestión que admitiría diversas respuestas, acordes a diversos criterios. No en vano, es sabido que todavía Miller, Galanter y Pribram (1960: 231) –valga un solo ejemplo– se presentaban como "conductistas subjetivos". Es evidente, por otra parte, que la paulatina frustración de algunas de las más osadas promesas explicativas del conductismo iría pareja al encandilamiento con los logros –a menudo, también, poco más que prometedores– de la investigación cognitiva que trasluce incluso en figuras tan proclives al escepticismo como la de Stich (1998: 649, *supra*). La insatisfacción con el conductismo, también la que pudiera tener un origen interno, no puede dejar de formar parte del recuento histórico –es elemental, dicho sea de paso, que lo mismo es válido *mutatis mutandis* de cara a la comprensión de las revoluciones políticas. Así, el concienzudo arqueo facturado por Mora (1992b, *supra*) de las dificultades internas que afrontaba el conductismo resulta tan plenamente pertinente como la anotación de Yela (1980/1996: 166) de que la desconfianza en su poder explicativo devasta también a quienes se habían contado entre sus más capaces espadas, tal como se refleja en las desalentadas reflexiones de Hilgard y Bower (1976):

> Se ha argumentado que las conductas más complejas, como el pensamiento y la solución de problemas, se entenderán más fácilmente una vez que se comprendan mejor las conductas simples en condiciones especialmente abreviadas […]. Después de treinta o cuarenta años sin avances notables en nuestra comprensión de la mente, tal argumento ha comenzado a sonar con tintineos engañosos. (Hilgard y Bower 1976: 465)

Con todo, entre los afluentes del cognitivismo y el funcionalismo puede hallarse, junto con los desengaños del conductismo, buena parte de sus conquistas. Hace falta –eso sí– aguzar la vista y advertir, de la mano de Rivière (1977), que:

> […l]a elegante regularidad de las leyes que nos brinda el análisis experimental del comportamiento sólo puede explicarse, a su vez, recurriendo a factores internos. Los resultados diferenciales de los diversos programas de refuerzo parecen indicar que el organismo, en definitiva, organiza […]. (Rivière 1977: 7)

Ese papel activo del organismo, que se convertiría en una divisa del cognitivismo y que éste, de acuerdo con la penetrante interpretación de Rivière, habría tenido que

aventurar precisamente para rendir cuenta de las regularidades desveladas bajo el programa conductista, es justo lo que ponían de relieve Bruner, Goodnow y Austin (1956) en su estudio sobre el proceso de formación de conceptos, cuyas líneas maestras consignaban en apenas tres preguntas cargadas de verbos de acción –que ellos mismos resaltaban:

> How do people *achieve* the information necessary for isolating and learning a concept? How do they *retain* the information gained from encounters with possibly relevant events so that it may be useful later? How is retained information *transformed* so that it may be rendered useful for testing an hypothesis still unborn at the moment of first encountering new information? (Bruner, Goodnow y Austin 1956: 51)

Menos afín que Campos-Roldán (1999) o Estany (1999) a la idea de que el cognitivismo constituyera una fuerza externa al conductismo, Leahey (2005: 395) alude, además de a los simposios de 1948 y 1956 y al empuje de la Alfred P. Sloan Foundation, a las reuniones del Grupo de Estudios de la Conducta Verbal que confluirían, junto con la actividad del Comité de Lingüística y Psicología del Consejo de Investigación en Ciencias Sociales y la que la Oficina de Investigación de la Marina venía desarrollando desde el final de la Segunda Guerra Mundial, en la creación del *Journal of Verbal Learning and Verbal Behavior*, en 1962 –hoy *Journal of Memory and Language*[45]. Como apunta Leahey, apelando al testimonio de James J. Jenkins, en los primeros tiempos de aquellos grupos de investigación las teorías mediacionales del conductismo tardío y las gramáticas al uso se veían como "variaciones de una misma línea de pensamiento" (Jenkins 1968: 539), pero poco a poco se iría entendiendo con claridad que los argumentos de Chomsky habían "[…] dinamitado la estructura [de la psicolingüística mediacional]". Si bien Leahey no parece tomar nota de ello –lo que apunta es, antes al contrario, que los primeros psicólogos cognitivos se limitaron a releer libremente la fórmula *E-r-e-R* (*E*stímulo – *r*espuesta mediacional – *e*stímulo mediacional – *R*espuesta), que habían aprendido de Hull, como entrada (E), procesamiento de información (r-e) y salida (R)–, los recuerdos de Jenkins parecen respaldar más bien una tesis de discontinuidad entre un conductismo que queda arrasado por las críticas chomskianas y un incipiente cognitivismo[46]. No en vano, cuando Dixon y Horton (1968: 580 *apud* Mandler 2002:

[45] Un minucioso estudio de la actividad del grupo, a cuyos simposios de 1959 y 1961 en Nueva York y 1966 en Lexington, Kentucky, acudió el mismo, se presenta en Mandler (2002). Aún en la reunión de 1961 –recuerda Mandler– uno de los participantes apuntaba que las posiciones debatidas podían desgajarse claramente entre las nítidamente afines a una psicología de estímulo y respuesta, y las que a falta de una denominación más perspicua acabó llamando "[…] 'non-S-R, or should it be anti-S-R?" (Cofer y Musgrave 1963: 374 *apud* Mandler 2002: 349). Ya Spence (1951), en un influyente análisis de "La interpretación teórica del aprendizaje", había trazado una distinción entre teorías asociacionistas, o S-R y teorías cognitivas, o S-S.

[46] Aunque, como se ha dejado anotado, los protagonistas fueran a menudo los mismos a ambos lados de la brecha. Así lo recalca Neisser (1967: 5), no sin cierto deje sarcástico:

350) trataron de recapitular los términos a los que había conducido el debate en los once años que mediaban entre la primera reunión del Grupo de Estudios de la Conducta Verbal, en 1955 en Minessota, y la más reciente, en 1966 en Lexington, su veredicto, aunque tentativo, no podía haber sido más claro: "[…] it appears that a revolution is certainly on the making".

El papel de los intereses bélicos en el desahucio del conductismo es puesto de relieve también por Mandler (2002: 343), quien, además de registrar el papel de la Oficina de Investigación de la Marina, apunta cómo las investigaciones psicoacústicas sobre el ruido desarrolladas en Harvard y en el MIT con vistas a su aplicación militar supusieron una "[…] desviación temprana del dogma conductista" que cristalizó, en 1951, en la fundación del Lincoln Laboratory del MIT, un vigorosísimo foco de investigación en el ámbito de la teoría de la detección de señales, cuyo ámbito de aplicación pronto desbordó el de la percepción para impregnar también el estudio de la memoria. Otro poderoso foco de posible rendimiento militar venía dado por el diseño de sistemas capaces de seguir por si mismos un objetivo –léase: un blanco–, o de guiarse en un entorno complejo: por aportar apenas un par de ejemplos, quizá no esté de más señalar, de la mano de Cordeschi (2002: 5), que ya en septiembre de 1915 el reportaje de portada de la revista popular *Electrical Experimenter* –dedicado a "The Wireless Torpedo"– mencionaba el interés de varios gobiernos hacia el robot fototrópico de J. Hammond Jr. y B.F. Miessner (*cf. infra*), de cuya primera exhibición pública había ofrecido la crónica sólo tres meses antes, y que hacía presagiar la creación de "armas inteligentes" capaces tal vez de decidir "[…] the present titanic struggle for the supremacy of Europe"; varias décadas, un tratado de paz y una nueva conflagración después, una de las primeras máquinas diseñadas con el ánimo de reproducir comportamientos inteligentes y, expresamente, de contribuir de esa manera a su comprensión –un pequeño ingenio construido por Richard A. Wallace, capaz de aprender a recorrer cierta clase de laberintos–, se presentó en el Congreso de la *Association for Computing Machinery* que se celebró en Pittsburgh en 1952 y que presidía C.V.L. Smith, de la Oficina de Investigación de la Marina[47]. Como se examinará con detenimiento, el proyecto de construir máquinas inteligentes es una de las fuerzas que con más brío impulsaron al cognitivismo, aunque, también, uno de los casos en que más nítidamente la nueva concepción de lo psicológico se hallaba ya prefigurada en los tiempos del conductismo.

> Today, happily, […] little or no defence [againt the behaviorist position] is necessary. Indeed, stimulus-response theorists themselves are inventing hypothetical mechanisms with vigor and enthusiasm and only faint twinges of conscience.

[47] Hay un minuciosísimo estudio sobre el papel de la investigación militar en el desarrollo de la simulación mecánica de procesos cognitivos en Cordeschi y Tamburrini (2006). El papel de la Oficina de Investigación de la Marina seguiría siendo decisivo en el desarrollo del cognitivismo más adelante: con su apoyo, por ejemplo, pudieron llevarse a cabo las investigaciones sobre heurísticas, sesgos, valores y marcos con las que Tversky y Kahneman (1974) y Kahneman y Tversky (1984) convulsionaron la teoría de la decisión racional.

A la hora de aquilatar la radicalidad de los cambios que condujeron del conductismo al cognitivismo, Estany (1999) registra también la reticencia de Thagard (1992) a tomar a uno u otro como paradigmas o –ni que decir tiene– como teorías. A diferencia de las teorías, que pugnan en el terreno de la capacidad explicativa, los *enfoques* –como conductismo y cognitivismo– lo hacen a su entender en virtud de su fertilidad en la producción de teorías –como los *proyectos de investigación* en la visión de la ciencia desplegada por Lakatos. Ésa sería, entonces, la razón fundamental por la que el conductismo cedió paso al cognitivismo: que el enfoque cognitivista resultaba más prometedor con vistas a la gestación de teorías convincentes. Como ha quedado dicho, sería la prosperidad del proyecto cognitivista lo que iría poco a poco minando la autoridad del conductismo para dictaminar aptitud para la investigación psicológica. Ahora bien, como recuerda Estany (1999: 194), "[...s]egún Thagard, el cognitivismo absorbió –más que eliminó– muchos de los conceptos del conductismo, por lo cual […], desde el punto de vista conductual, tiene un 'carácter de revolución blanda'". Acaso más desencantado se muestra Leahey, quien tras conducir su análisis de los elementos de continuidad entre el conductismo mediacional y el cognitivismo a la conclusión de que "[…] la psicología cognitiva […] se entiende mejor como la última forma del comportamentalismo, con importantes afinidades con algunas formas históricas de conductismo" (Leahey 2005: 396), cierra su estudio de los orígenes del cognitivismo dejando bosquejada la idea de que, como suele ocurrir también con las revoluciones políticas, "[…] la revolución cognitiva no fue más que una ilusión" (Leahey 2005: 397).

La paridad metodológica entre cognitivismo y conductismo –la pervivencia del "utillaje metodológico empleado por los conductistas", en palabras de Mora (1992b: 108)– queda particularmente recalcada en el estudio sobre la figura de Watson de Morris y Todd (1999), con la escueta mención de que:

> Methodologically, psychology is little changed today: Psychologists still study behavior, but no longer as a subject matter in its own right, as it was for Watson. Instead, behavior is a basis for objective inferences about brain, mind, and cognition and theories thereof […] (Morris y Todd 1999: 16)

Que esa diferencia sea tan nimia como parece que se quisiera hacer ver, sin embargo, dista de estar claro. Repárese –por ejemplo– en que una lectura del giro "estudiar la conducta" cuya superficialidad no es más reprobable que la del argumento de Morris y Todd desdibuja también el contraste entre la "psicología fisiológica" de Wilhelm Wunt y el análisis experimental de la conducta de Watson: también Wundt estudiaba la conducta –los informes introspectivos de los sujetos del laboratorio de Leipzig, es decir, un fragmento de su conducta verbal– como fundamento de inferencias objetivas que incorporaría a sus teorías psicológicas o, más esporádicamente, fisiológicas. No en vano, así lo acentuaba ya Angell (1913) al tratar –en el sentido opuesto– de poner coto al ánimo de prescindir de la introspección que adivinaba en el desarrollo inminente de la psicología:

> But we must not forget that whenever we avail ourselves of language as a mode of approach to objective behavior, we are apt to find ourselves compromising with introspection, for such language may simply report mental states in the ordinary introspective fashion. (Angell 1913: 263)

Es obvio, ahora bien, que una interpretación de "estudiar la conducta" que establece continuidad metodológica tanto entre conductismo y cognitivismo como entre introspeccionismo y conductismo es forzosamente una interpretación demasiado vaga. Por otro lado, es indudable que un análisis pormenorizado de los métodos de investigación empleados de hecho bajo la tutela del conductismo y bajo la del cognitivismo revelaría tan significativas diferencias como semejanzas: valga tan sólo apuntar, como uno de los epígrafes seguros de ese estudio, el papel primordial de la simulación computacional en el repertorio metodológico cognitivista –pese a las airadas protestas de Skinner (1985).

No conviene por lo demás, sea como fuere, a una defensa de la continuidad de fondo entre conductismo y cognitivismo, ni menos aún a la del valor de determinadas intuiciones que germinaron en el seno del movimiento conductista, un excesivo hincapié en la vigencia de las restricciones metodológicas, o de las prácticas higiénicas (*supra*), que éste impulsara. No conviene porque –como ha sabido recalcar Campos-Roldán (1999)–, por afianzada que quede, incluso en los manuales, la tesis de que "[…l]a piedra fundamental, la metodología conductista, ha resistido obstinadamente, y en la actualidad se la debe considerar como una contribución sólida y evidentemente duradera" (Marx y Hillix 1963/1979: 195), no lo estará menos su contrapunto:

> Sin embargo, una piedra no es un edificio, y una restricción metodológica no es un sistema; de modo que […] tampoco hay en la actualidad un sistema *completo* que se denomine *conductismo*. (Marx y Hillix 1963/1979: 195)

A juicio de Estany (1999), pese a todo, la reelaboración de la noción de paradigma emprendida por Lachman, Lachman y Butterfield (1979) permite reconocer ese núcleo metodológico que el cognitivismo toma del neoconductismo sin vulnerar la tesis de inconmensurabilidad ni, por tanto, la de que uno y otro constituyen diferentes paradigmas. El concepto de representación mental sería, para ellos, parte de la *idea preteórica* central del cognitivismo: "[…] que cuando los seres humanos realizan funciones cognitivas actúan como un sistema de procesamiento de información" (Estany 1999: 186) –idea preteórica sería también, por ejemplo, la de universo armónico en Kepler. Ciertamente, la necesidad de postular representaciones internas fue uno de los motivos recurrentes de las primeras reivindicaciones cognitivistas casi como la inutilidad de la introspección lo había sido en los textos fundacionales del conductismo; por las mismas razones, sin duda, el carácter a su juicio enteramente superfluo –es decir, extravagante– de tal postulación se convertiría en uno de los caballos de batalla en los tardíos pleitos de Skinner (1985,

1987, 1989, 1990) contra lo que él entendía como un nocivo retorno del mentalismo. Reparar el valor de los conceptos con carga intencional –pensamientos, proposiciones, esquemas, guiones, imágenes, deseos… en fin, representaciones internas– que la psicología introspectiva había tomado como lenguaje observacional fundamentándolos sobre "[…] observaciones establecidas en términos estrictamente *extensionales*" se perfila a ojos de Rivière (1991b: 134) como la contribución decisiva del cognitivismo a nuestra comprensión de lo mental. Si bien es cierto que tal contribución sólo pudo darse tras haber quedado asimilada la tentativa conductista de construir "[…] una psicología escuetamente extensional, tanto en su lenguaje observacional como en el teórico", no lo es menos que requirió, también, imbricar aquellas longevas nociones en una estructura teórica fresca, la que, como se estudiará con cierto detenmiento, proporcionaban los vivificantes avances de la teoría de autómatas y el análisis del concepto de computación, "[…] que implican el compromiso con un mecanicismo formal y abstracto" (Rivière 1991b: 129).

En uno de los trabajos fundacionales del cognitivismo, el estudio sobre *Planes y Estructura de la Conducta* de Miller, Galanter y Pribram (1960), el carácter preteórico del concepto de representación mental al que aluden Lachman, Lachman y Butterfield (1979) quedaba reflejado, quizá cándidamente, en el modo en que se plantea el desacuerdo fundamental entre "teóricos del reflejo" y cognitivistas. Por un lado –dicen Miller, Galanter y Pribram (1960: 16-17)–, tenemos a "los optimistas, que pretenden descubrir la dependencia [entre la conducta del organismo y lo que sucede en su entorno] sencilla y directamente […] según el patrón fisiológico clásico del arco reflejo"; frente a ellos estarían "los pesimistas, que piensan que los organismos vivos son complicados, tortuosos, mal diseñados para los fines de la investigación, y cosas por el estilo". Sería entonces ese cierto talante lo que con anterioridad a cualquier formulación teórica, a cualquier conjetura o acúmulo de observaciones, habría llevado a los cognitivistas a la convicción de que:

> […] el efecto que tendrá un acontecimiento sobre la conducta depende de cómo se represente dicho acontecimiento en la representación que el organismo tiene de sí mismo y de su universo. Están completamente seguros de que todas las correlaciones entre estímulo y respuesta deben ser mediatizadas por una representación organizada del entorno, un sistema de conceptos y relaciones dentro del cual se encuentra el organismo. Un ser humano, y de igual manera probablemente otros animales, elabora una representación interna, un modelo del universo, un simulacro, un mapa cognitivo, una *imagen*. (Miller, Galanter y Pribram 1960: 17)

Era previsible que tal compromiso con la idea de que el organismo albergue una representación interna de su entorno –y de sí– despertara las iras de quien, como Skinner, lo había reprobado ya en la figura de Tolman. La tesis cognitivista que Skinner procuraría reiteradamente escarnecer –de entrada, bautizándola como "teoría de la copia"– es la de que, tal como la destila Ringen (1990):

> Intelligent human action is guided by contentful cognitive states that to a greater or lesser degree succeed or fail in representing states of affairs in the world in which the actions guided by these states occur. Mental representations cause behavior. (Ringen 1990: 171)

La eficacia causal de los estados mentales, así pues, aparece ligada tan estrechamente a su función adaptativa –guiar al organismo en el entorno que habita– como a su capacidad de representar éste *verídica o erróneamente*, es decir, al principio normativo que ineludiblemente se le impone: de nuevo, la posibilidad del error. El propio Ringen (1990: 174) toma nota de las reflexiones de Charles Taylor (1964) al respecto, en las que el concepto de entorno intencional prefigura los *mundos nocionales* ensayados por Dennett (1982):

> In an "intentional system" […] the condition of an action occurring is that it be believed to be adequate to the goal, and not simply that it is, in fact, adequate. And, the two may not go together. The situation as it really is may differ from the situation under its intentional description for the agent, that is, the intentional description may not hold of it […]. The teleological account holds not of the agent in its "geographical" environment, but of the agent in its "intentional" environment, the environment as it is for him. (Taylor 1964: 62)

A juicio de Skinner, por supuesto, el entorno intencional no era sino una fantasía falaz que habría de torcer –si nos dejábamos hechizar por ella– la cabal investigación de las relaciones funcionales entre estímulos y respuestas, en la que la adaptación del organismo a su entorno quedaría felizmente descrita. Así vistas –en contraste con las de Skinner– las certidumbres preteóricas del cognitivista no parecen, en efecto, muy diferentes de las que en su día llevaran al estoico a asegurar que "lo que turba a los hombres no son los sucesos, sino las opiniones acerca de los sucesos" (*Enquiridión*, V) –lo hace Epicteto de Frigia a tenor de su fe en que si la muerte fuera en sí misma algo terrible, así se lo habría parecido a Sócrates. Cómo se forman esas opiniones, y cómo exactamente rigen lo que nos turba y también lo que hacemos son las preguntas que constelan la investigación cognitivista. De la orilla de la acción, la sinopsis del proyecto quedó ya trazada con insuperable concisión por Miller, Galanter y Pribram (1960: 23): "El problema consiste en describir la forma en la que la representación interna que un organismo posee de su universo controla las acciones".

Fingida austeridad, o entender qué aprendemos

Entre los primeros argumentos esgrimidos en pro de la teorización sobre representaciones internas cobraron un especial relieve los de Chomsky (1959) y los de Fodor (1968). Se trata, en realidad, de líneas de razonamiento parejas: la observación medular en las tesis de Chomsky hacia 1959 era la de que la pobreza de los estímulos lingüísticos que recibe un niño impediría que éste aprendiera su lengua materna a no ser que contara con un sistema innato de reglas y representaciones; en el caso de Fodor (1968), lo que nos obligaría de recurrir a reglas y representaciones

mentales es la necesidad de explicar determinados efectos de transposición en el aprendizaje –cuya estrecha relación con los fenómenos de constancia perceptiva ya barruntaba Köhler (1917)– para los que la información presente en los estímulos parecía rotundamente insuficiente. Así, por ejemplo, Fodor se muestra firmemente convencido de que:

> Cualquier intento serio de construir una teoría de la percepción psicológicamente viable tendría que dar cuenta […] del hecho de que la práctica se generaliza a menudo a objetos relacionados sólo de forma muy abstracta con el objeto directo de esa práctica. Y no es fácil imaginar cómo se podría llevar esto a cabo sin asumir que el concepto que se tenga de una cara, o de una melodía o de una forma (es decir, "las recetas para reconocer" formas, melodías y caras) incluya la representación de la estructura formal de cada uno de estos dominios, y sin asumir igualmente que el acto de reconocimiento consiste en la aplicación de esa información a la integración de los distintos *inputs* sensoriales que actúan en cada momento. (Fodor 1968: 55)

La advertencia de Fodor tiene lugar en el contexto de una viva discusión sobre la idea de destreza que parece traslucir en el estudio de Ryle (1949) de determinadas capacidades tradicionalmente investidas de tintes psicológicos. La cuestión debatida es si cabe, por ejemplo, ofrecer un análisis netamente disposicional de la capacidad de interpretar una melodía –un caso que, como quedó anotado de mano de Place (1999: 373, *supra*), parece hondamente enraizado en el pensamiento de Ryle, y podría provenir de una observación de Wittgenstein (1958)[48]. Al cabo de una laboriosa disputa con las posiciones expresadas por Ryle y las que podrían derivarse de ellas, Fodor esboza la conclusión decididamente cognitivista de que:

> En resumen, si lo que tienen en común las distintas formas de interpretar "Lillibullero" es algo de carácter abstracto, se sigue claramente que el sistema de expectativas, que constituye la receta que uno tiene para oír la canción, debe ser igualmente abstracto. […S]e deberá admitir igualmente que el aprender a reconocer melodías lleva consigo internalizar y aplicar conceptos complejos –seguramente como resultado de operaciones mentales de análoga complejidad. (Fodor 1968: 56-57)

Ahora bien: aunque Köhler hubiera comprendido que en ciertos aspectos la transposición de destrezas aprendidas y la percepción estable del entorno que nos proporcionan las constancias perceptivas han de tener facetas comunes, su lealtad a una concepción isomorfista de las relaciones entre –como él mismo diría– el campo mental y el campo cerebral obstaculizó el desarrollo de una teoría detallada de los procesos internos que pudieran sustentar tales capacidades. A la par que delataba las limitaciones del modelo de aprendizaje de secuencias motoras complejas bosquejado por Watson, Lashley (1951) denunciaría que la teoría de Köhler ni siquiera intentaba abordar la cuestión que desde su perspectiva resultaba más acuciante:

[48] Cuando no, acaso, del *tópos* del músico, cuyos ecos conducen, como se ha visto *supra*, hasta Whytt, Stahl, von Haller o de la Mettrie.

> El problema neurológico consiste en gran medida, si no completamente, en la traducción del patrón aferente de estímulos en el patrón eferente. La teoría del campo no incluye, en su formulación actual, ninguna indicación acerca de la forma en que las fuerzas del campo inducen y controlan el patrón de la actividad eferente. Se aplica a la experiencia perceptiva pero parece acabar ahí. (Lashley 1951: 230 *apud* Miller, Galanter y Pribram 1960: 21).

Por supuesto, si Miller, Galanter y Pribram se hacían eco de las críticas de Lashley a los gestaltistas –a quienes ellos acogen como "otros teóricos cognitivos" (Miller, Galanter y Pribram 1960: 20)– era para fundamentar su propia apuesta por una explicación de la conducta en términos de imágenes y planes, es decir, de representaciones internas. El trayecto que preocupa a Miller, Galanter y Pribram parte del estado cognitivo, de la creencia, la hipótesis o la abstracción, para dirigirse a la conducta. Lo que pretenden salvar, como ellos mismos apuntan, es el vacío que queda entre el mapa cognitivo que, según había hipotetizado Tolman, permite a la rata saber dónde se encuentra la comida y los pormenores de la conducta apetitiva de la rata. Una sagaz objeción de Guthrie contra Tolman –que servía de inspiración a Miller, Galanter y Pribram, y que también evocaría Pylyshyn (1984: 79)– hace casi palpable ese vacío: "En lo que toca a la teoría, la rata ha quedado sumida en sus pensamientos; si al final llega hasta la caja de comida, esto es cosa suya, no es cosa que interese a la teoría" (Guthrie 1953: 172 *apud* Miller, Galanter y Pribram 1960: 19). Pues bien, la otra orilla de ese mismo trayecto, en la que arriba al estado cognitivo la información proveniente del estímulo, es en la que había detenido su mirada Fodor (1968). A ojos de Köhler, en cambio, ambos espacios parecían quedar cubiertos por la mera mención de la tesis de isomorfismo; a ojos de Ryle, ni siquiera parecían existir. La tarea primera del cognitivismo sería señalarlos, hacer patente la necesidad de cartografiarlos. No es descabellada, pues, la caracterización del cognitivismo como una peculiar suerte de pesimismo, siendo el empeño del cognitivista avisar de la existencia de unas penumbras que otros niegan, o ante las que se dan por satisfechos[49]. En ese sentido, la posición de Köhler no era demasiado diferente de la que sería la del propio Ryle –y que se puede rastrear también en muchas críticas al cognitivismo de inspiración wittgensteiniana–: la destreza motora o perceptiva, como la disposición en Price (1953), acaba apareciendo como un *datum*, un lecho rocoso (*cf.* Wittgenstein 1953: §217; 1969: *passim*) ante el cual ha de detenerse la prospección[50].

[49] A la misma idea se diría que apunta Pylyshyn (1984: 13) cuando, al hilo del trabajo de Hochberg (1968), describe el empeño de traducir las leyes gestálticas a un lenguaje fisicalista como fruto de un "optimismo prematuro". Sólo articular un conjunto de propiedades físicas del estímulo que pueda tomarse como *analysans* de su complejidad percibida se revelaría ya, según el diagnóstico de Pylyshyn, como una tarea inacabable. También Mandler (2002: 346) menciona a Hochberg como uno de los pioneros de la investigación en procesos cognitivos en percepción.

[50] En su acerada crítica del modo en que Locke proponía entender la formación de conceptos, Bennett (1971: 15), sin embargo, recurre a Wittgenstein como antídoto a una terminación prematura del esfuerzo explicativo, al concluir que "Locke probably overlooked the need to classify ideas because they are in the mind and, as Wittgenstein remarks, 'the mind' often serves as a haven for the not-to-be-explained".

La cercanía de Wittgenstein es perspicua, por ejemplo, en los razonamientos que tardíamente emplea Malcolm (1971) –que había asistido a las clases sobre fundamentos de las matemáticas que Wittgenstein impartiera en Cambridge durante el curso de 1938/1939– con el propósito de desarticular el uso que da el cognitivismo, cuando postula procesos y estructuras cognitivas, a las nociones de *proceso* y *estructura*. Si bien el lenguaje ordinario –atestigua Malcolm– describe en ocasiones cómo alguien atraviesa el proceso de recordar algo, a lo que se refiere es en realidad al proceso de *intentar* recordar algo: reconstruir deliberada, conscientemente, una secuencia de acontecimientos pasados a fin de –cabría decir– elaborar el recuerdo que nos elude. Pero es ilegítimo inferir de esto que siempre que uno recuerda algo, incluso cuando el recuerdo nos viene dado espontáneamente y sin esfuerzo deliberado por nuestra parte, se produzca un proceso cognitivo, el de recordar. Dado que, a diferencia de los procesos psicológicos que detalla el discurso de sentido común, los presuntos procesos cognitivos no se presentan a la consciencia, esa inferencia indebida nos aboca irremediablemente, además, a darle a los presuntos procesos cognitivos –recordar, entender, reconocer, etc.– una naturaleza oculta, dotándolos del aura de la profundidad[51]. Entonces –asegura Malcom con el reconocible tono clínico del conductismo lógico:

> We feel that when a person recognizes something, in addition to the various manifestations or characteristics accompaniments of recognition, something must go on inside. This is the "inner process" of recognition. (Malcolm 1971: 387)

Pero el conductismo es el antídoto para esta tentación. Aunque –pongamos por caso– reconocer a alguien no pueda ya equipararse a las propias expresiones de reconocimiento, o a la disposición a mostrarlas –algo temperada está en 1971 la audacia de los primeros tiempos del conductismo, de la que había hecho alarde el propio Malcom (1951, *infra*) en su análisis de los sueños como conducta de vigilia–,

[51] Acerca de la naturalidad con la que Malcolm parece asumir la existencia de procesos psicológicos siempre y cuando estos resulten asequibles a la introspección, y al mismo tiempo esquivar la cuestión de que puedan existir procesos psicológicos que no lo sean –como, ciertamente, los que acostumbran a postular los psicólogos cognitivos–, *cf*. Martin (1973).

Si tenemos en cuenta que la irrelevancia de la distinción entre lo consciente y lo inconsciente de cara a la captura de (cuando menos algunas, modestas) generalizaciones relevantes acerca de la conducta humana ha llegado a ser planteada como uno de los descubrimientos cardinales del cognitivismo, la *petitio* en que parece incurrir Malcolm a este respecto se torna más flagrante. Así, por ejemplo, Pylyshyn (1984: 265) escribe:

> I consider it an empirical discovery of no small importance that if we draw the boundary around phenomena in such a way as to cut across the conscious-unconscious distinction, we find that we can at least formulate moderately successful minitheories in the resulting cluster.

En qué medida tal descubrimiento viniera prefigurado por el trabajo de Sigmund Freud es cuestión controvertida. Un lugar desde el que adentrarse en ese debate puede encontrarse en Erdelyi (1985).

queda un vastísimo refugio para evitar lo interno, un refugio tan ancho como el mundo:

> Recognizing someone is not an act or process, over and above, or behind, the expression of recognition in behavior. But also, of course, it is not that behavior. […] [I]t is the facts, the circumstances surrounding that behavior, that give it the property of expressing recognition. This property is not due to something that goes on inside. (Malcolm 1971: 387)

Bien podría ser, desde luego, que para distinguir el reconocimiento, el recuerdo o la comprensión genuina de cualquier posible impostor nos baste tal o cual conjunto de propiedades estrictamente físicas del entorno –entiéndase: propiedades susceptibles de quedar identificadas, no sólo como particulares sino también en términos de su inclusión en clases (a efectos de su participación en generalizaciones empíricas) mediante un aparato conceptual desprovisto de toda alusión, abierta o velada, a estados internos del organismo. Bien podría ser –parece que debería conceder Malcolm– también lo contrario. En el *experimentum crucis* imaginario elegido por Malcolm, la balanza parece inclinarse a su favor: podemos diferenciar al hombre que ha reconocido a alguien del que finge reconocer a la última de cada diez personas con las que se cruza, y que resulta cruzarse con alguien a quien "nunca ha visto" (Malcolm 1971: 387); no parece que para hacerlo nos sea preciso mencionar nada que "ocurra en el interior" de uno o de otro. Pero no parece que se haya puesto el mismo afán en imaginar un caso adverso: cómo diferenciar, supongamos, a un hombre que reconoce a alguien de otro que finge reconocer a alguien a quien ha visto antes, pero a quien en realidad no reconoce –cómo hacerlo, esto es, aludiendo únicamente a propiedades estrictamente físicas del entorno[52].

Pero sea como sea, la discusión es irrelevante respecto a las conclusiones que el conductista trata de establecer. Incluso si pudiéramos diferenciar cada caso en que un sujeto reconoce (recuerda, comprende, cree, *etc*.) sin aludir sino a sus conductas y a las propiedades de su entorno, eso no dejaría sin sentido la pregunta acerca de cómo logra el sujeto desplegar esas notables capacidades –o por qué, en ocasiones triviales o desdichadas, deja de hacerlo. Que esa pregunta sea ilegítima, o que para responderla nos esté vedado teorizar acerca de procesos cognitivos –procesos análogos a los que se dan, por ejemplo, cuando nos esforzamos por recordar algo, pero de los que no siempre tengamos consciencia–, dista mucho de haber quedado establecido por el argumento de Malcolm. A decir verdad, tampoco Malcolm se muestra tan ambicioso: se conforma con señalar que:

[52] O peor: cómo hacer otro tanto cuando el verbo psicológico en cuestión no sea de índole epistémica, como "reconocer", sino doxástica, como "creer", y las circunstancias mundanas –quién ha visto a quién, admitiendo el confuso empleo de "ha visto" como descripción no psicólogica de tales circunstancias– queden inermes.

> […] if this point were understood by philosophers and psychologists, they would no longer have a motive for constructing theories and models for recognition, memory, thinking, problem solving, understanding, and other "cognitive processes". (Malcom 1971: 387)

Lo que parece eludir a Malcolm es que muchos filósofos y psicólogos encontraban sus motivos para construir tales teorías precisamente en la desconfianza respecto a lo que él da por hecho: que sea factible una descripción convincente de las regularidades observadas en la conducta que prescinda de toda mención de estados internos del sujeto de dicha conducta para atenerse sólo a la propia conducta y las circunstancias que la rodean[53]; que tal descripción, aun de resultar viable, constituyera *ipso facto* una explicación de dichas regularidades. Nada de ello –eso sí– hace menos valiosa una de las lecciones que cabe extraer de los razonamientos de Malcolm, como de los de los gestaltistas: que a la hora de construir una explicación psicológica no conviene menoscabar la riqueza del entorno[54], por mucho que advertir su pobreza haya sido, también, cardinal para nuestra comprensión de lo mental. Dicho de otro modo, que lo que ha sido para el cognitivismo una intuición fundacional –la de la pobreza del estímulo: *cf*. Chomsky (1959), Fodor (1968), *supra*– no fosilice como un dogma.

La prolija discusión con Ryle acometida por Fodor –y a la que Malcom, sin citar a Fodor, parece contestar– tendría no obstante una repercusión en los círculos de la psicología experimental mucho menor que la cosechada por la reseña de *Verbal Behavior* (Skinner 1957) que el editor de *Language*, Bernard Bloch, solicitó a un joven Noam Chomsky (1959), a la sazón profesor de Lingüística del Instituto de Tecnología del MIT, a cuyo claustro se incorporaría ese mismo año el propio Fodor. La reseña de Chomsky –ha apuntado Baars (1986: 338)– acabaría por ser, con creces, más influyente que el propio libro de Skinner.

A veces se ha pasado por alto el hecho de que, apenas iniciada la reseña, Chomsky establece el marco de objetivos y criterios explicativos que comparte con Skinner, y que ha de posibilitar el diálogo. Queda claro así que no está en litigio la idea de que el análisis funcional de la conducta, vagamente entendido como la identificación de las variables que controlan la conducta y de las relaciones entre ellas que determinan una respuesta en particular, constituye una meta razonable de la investigación. Lo que resulta "sorprendente", en cambio, es:

---

[53] Exactamente la suposición contraria, por ejemplo, a las conclusiones que –haciendo notar de paso que se trata de "[…] una cuestión empírica, no lógica"– adelanta Pylyshyn (1984: 12):

> Most nonbehaviorists believe that […] what people do depends to a great extent on what they believe at the moment, how they perceive a situation […], on what they think will be the consequences of various behaviors, and so on.

[54] La misma lección, por otra parte, que James J. Gibson siempre trató de inculcar a sus colegas cognitivistas. Una somera discusión del realismo perceptivo directo de Gibson, en términos muy cercanos a los de la presente lectura de Malcolm, puede encontrarse *infra* con relación a su valoración por parte de Pylyshyn (1984: 182-183).

> […] the particular limitations he [Skinner] has imposed on the way in which the observables of behavior are to be studied, and, above all, the particularly simple nature of the function which, he claims, describes the causation of behavior. (Chomsky 1959: 413)

Del cotejo de esas limitaciones de orden metodológico y la envergadura de las conclusiones de Skinner con las instancias concretas de explicación de conductas verbales con que él mismo guarnece su propuesta nace la doble acusación que vertebra el trabajo de Chomsky, a saber: el vocabulario teórico al que recurre Skinner en su explicación de la conducta verbal no es sinónimo, sino homónimo, del utilizado en sus estudios experimentales sobre condicionamiento operante; además, está plagado de vacuidades y de tintes mentalistas deliberadamente velados que lo hacen explicativamente inerte, cuando no introducen más oscuridad que claridad.

El envite de Chomsky es, desde el principio, arriesgado:

> The magnitude of the failure of this attempt to account for verbal behavior serves as a kind of measure of the importance of the factors omitted from consideration, and an indication of how little is really known about this remarkably complex phenomenon. (Chomsky 1959: 414)

La homonimia de las nociones de estímulo, respuesta, reforzamiento y condicionamiento tal como se utilizan en la insalvable concepción skinneriana de la conducta verbal y en la valiosa literatura experimental sobre condicionamiento operante es, en efecto, el núcleo de la argumentación de Chomsky. Como la reseña trata de hacer patente mediante la acumulación de citas de *Verbal Behavior*, "[…] the insights that have been achieved in the laboratories of the reinforcement theorist, though quite genuine, can be applied to complex human behavior only in the most gross and superficial way […]" (Chomsky 1959: 414). La consecuencia menos grave de esa endeble aplicación sería, de acuerdo con Chomsky, que los conceptos básicos del paradigma de condicionamento operante quedarían trocados, al trasladarse a la explicación skinneriana de la conducta verbal, en "[…] mere homonyms, with at most a vague similarity of meaning […]" (Chomsky 1959: 416) respecto al esquema original; su potencia explicativa menguaría entonces hasta la de vagas conjeturas, "[…] analogic guesses (formulated in terms of a metaphoric extension of the technical vocabulary of the laboratory) […]" (Chomsky 1959: *ibid.*). Aunque la maniobra logre forjar "[…] the illusion of a rigorous scientific theory with a very broad scope […]" (Chomsky 1959: *ibid*.), la quimera se desvanece ante un somero análisis crítico, que el propio Chomsky articula. Todo el esfuerzo de Skinner por esquivar la referencia a estados mentales en su vocabulario teórico se revela inútil, ya que si logra hacerlo es sólo a costa de infectar de veladas referencias a lo mental el vocabulario conductista. En las contundentes palabras de Chomsky:

> The way in which these terms are brought to bear on the actual data indicates that we must interpret them as mere paraphrases for the popular vocabulary commonly used to

> describe behavior and as having no particular connection with the homonymous expressions used in the description of laboratory experiments. Naturally, this terminological revision adds no objectivity to the familiar *mentalistic* mode of description. (Chomsky 1959: 420)

El ardid pergeñado por Skinner puede tener, no obstante, consecuencias peores: una cosa es verlo como un anodino artefacto léxico, y otra bien distinta advertir de que su empleo dificulte la comprensión de aquello que pretende explicar. Pero es precisamente en esta dirección en la que a lo largo de su evaluación del trabajo de Skinner bascula el dictamen de Chomsky. Ya antes de desplegar los detalles de su análisis, Chomsky adelanta que:

> […] with a literal reading (where the terms of the descriptive system have something like the technical meanings given in Skinner's definitions) the book covers almost no aspect of linguistic behavior, and […] with a metaphoric reading, it is no more scientific than the traditional approaches to this subject matter, *and rarely as clear and careful*. (Chomsky 1959: 416, énfasis añadido)

Tras el análisis, la conclusión quedará refrendada casi con las mismas palabras, pero la acusación de pérdida de claridad se verá entonces sustanciada:

> […] if we take his [Skinner's] terms in their literal meaning, the description covers almost no aspect of verbal behavior, and if we take them metaphorically, the description offers no improvement over various traditional formulations. The terms borrowed from experimental psychology simply lose their objective meaning with this extension, and take over the full vagueness of ordinary language. Since Skinner limits himself to such a small set of terms for paraphrase, many important distinctions are obscured. (Chomsky 1959: 432)

Como se ha anunciado, el amaño impugnado por Chomsky concierne a la carga mentalista emboscada en el uso que hace Skinner de conceptos de fingida austeridad conductista. Sobre las nociones de reforzamiento y condicionamiento, en particular, lo que pesa es fundamentalmente una acusación de vacuidad. Así, por ejemplo, la revisión pormenorizada del uso que Skinner da al concepto de reforzamiento en la explicación de conductas concretas haría imposible, según Chomsky, dar significado a las tesis generales en las que aparece dicho concepto:

> Skinner does make it very clear that in his view reinforcement is a necessary condition for language learning and for the continued availability of linguistic responses in the adult. […] However, the looseness of the term reinforcement as Skinner uses it in the book under review makes it entirely pointless to inquire into the truth or falsity of this claim. Examining the instances of what Skinner calls reinforcement, we find that not even the requirement that a reinforcer be an identifiable stimulus is taken seriously. In fact, the term is used in such a way that the assertion that reinforcement is necessary for learning and continued availability of behavior is likewise empty. (Chomsky 1959: 420)

El veredicto de Chomsky se reitera entonces con toda su dureza: tal utilización del concepto de reforzamiento no sólo no introduce objetividad alguna en la explicación, sino que la enturbia.

> The phrase "X is reinforced by Y (stimulus, state of affairs, event, etc.)" is being used as a cover term for "X wants Y," "X likes Y," "X wishes that Y were the case," etc. Invoking the term reinforcement has no explanatory force, and any idea that this paraphrase introduces any new clarity or objectivity into the description of wishing, liking, etc., is a serious delusion. The only effect is to obscure the important differences among the notions being paraphrased. (Chomsky 1959: 421)

En una fugitiva incursión en el ámbito de la sociología de la ciencia, Chomsky (1959: 421) anota en ese mismo párrafo que "[…] we can only conclude that the term reinforcement has a purely ritual function"; la observación –qué duda cabe– rayaría el insulto a ojos de Skinner. Tanto peor, después de un examen algo más escueto pero perfectamente análogo, el uso del concepto de condicionamiento por parte de Skinner (1957) es objeto de idéntica censura, salvo en que el ritual se ha convertido ya en mera farsa: "[…t]o speak of 'conditioning' or 'bringing previously available behavior under control of a new stimulus' in such a case is just a kind of play-acting at science" (Chomsky 1959: 422). El inclemente retrato del conductismo que emanaba de la lectura de Chomsky era al fin y al cabo –como después señalaría Adárraga (1991: 40)– el de "[…] una especie de 'mentalismo avergonzado', disfrazado mediante una jerga, puramente ritual, de estímulos y respuestas".

Con igual severidad consigna Chomsky, al lado de la vacuidad de los conceptos de reforzamiento y condicionamiento según se emplean en *Verbal Behavior*, la transfiguración a la que Skinner somete a las nociones de estímulo y respuesta –con respecto, se entiende, a su uso en la literatura experimental. Tan penetrante es dicha metamorfosis que, a su término, los estímulos han dejado de ser siquiera acontecimientos ambientales, sobrecargados como quedan de implicaciones mentalistas:

> Stimuli are no longer part of the outside physical world; they are driven back into the organism. […T]alk of stimulus control simply disguises a complete retreat to mentalistic psychology. (Chomsky 1959: 417)

Lo mismo vale, *mutatis mutandis*, para la noción de respuesta. Particularmente acre se muestra Chomsky con la forma en que Skinner habla de las variaciones en la probabilidad de emisión de una respuesta, cajón de sastre en el que cabe buena parte del aparato explicativo de la psicología de sentido común. La acusación de haber ritualizado el uso del vocabulario científico quedaba ya esbozada en estas líneas: entre afirmar que el sujeto tiene, por ejemplo, interés en hacer *C* y que existe una mayor probabilidad de que emita la respuesta *C*, lo único que sustenta la inclinación de Skinner por la segunda variante es su voluntad de usurpar la fisonomía de una explicación científica acabada.

> It is not unfair, I believe, to conclude from Skinner's discussion of response strength, the basic datum in functional analysis, that his extrapolation of the notion of probability can best be interpreted as, in effect, nothing more than a decision to use the word probability, with its favorable connotations of objectivity, as a cover term to paraphrase such low-status words as interest, intention, belief, and the like. (Chomsky 1959: 419)

La argumentación articulada por Chomsky resultaría aún más concluyente –acaso por razones estéticas– al acentuarse la forma de dilema en que naturalmente se clausuraba: un dilema, como anotarían Miller, Galanter y Pribram (1960: 33), entre la irremediable ambigüedad de los conceptos conductistas al rebasar los límites del laboratorio de condicionamiento, y la irrelevancia de esos mismos conceptos, de cara a la comprensión de extensas regiones del comportamiento humano, cuando quedaban uncidos a la estricta interpretación del análisis funcional.

> Questions of this sort pose something of a dilemma for the experimental psychologist. If he accepts the broad definitions, characterizing any physical event impinging on the organism as a stimulus and any part of the organism's behavior as a response, he must conclude that behavior has not been demonstrated to be lawful. In the present state of our knowledge, we must attribute an overwhelming influence on actual behavior to ill-defined factors of attention, set, volition, and caprice. If we accept the narrower definitions, then behavior is lawful by definition (if it consists of responses); but this fact is of limited significance, since most of what the animal does will simply not be considered behavior. Hence, the psychologist either must admit that behavior is not lawful (or that he cannot at present show that it is –not at all a damaging admission for a developing science), or must restrict his attention to those highly limited areas in which it is lawful (e.g., with adequate controls, bar-pressing in rats; lawfulness of the observed behavior provides, for Skinner, an implicit definition of a good experiment). (Chomsky 1959: 416)

Habrá de ir volviéndose más claro, en el transcurso de esta investigación, que el origen último de la dificultad advertida por Chomsky es la decisión que tempranamente adoptase Skinner de que los conceptos de estímulo y respuesta no quedaran sometidos a definiciones categóricas, en términos físicos, sino a definiciones funcionales, de orden disposicional: es decir, por ejemplo, la decisión de contar como instancias de un determinado tipo de respuesta cualquier aspecto de la conducta cuya probabilidad de emisión viniera ligada a la presencia de cierto estímulo –digamos, el reforzador–, estímulo, por lo demás, cuyas características físicas tampoco tenían por qué permanecer inmutables mientras no se alterara su efecto sobre la respuesta. De ahí, dicho sea de paso, que Skinner ni siquiera tuviera costumbre de observar a los inquilinos de su animalario, contentándose con revisar los registros acumulativos de sus respuestas –como no podía ser de otro modo toda vez que, por ejemplo, en *Schedules of Reinforcement* (Ferster y Skinner, 1957) se analizan cerca de 70.000 horas de registros conductuales. En todo caso, se argumentará más adelante que la adopción de un vocabulario físico o de un vocabulario funcional en la caracterización de estímulos y respuestas es todavía un

asunto irresuelto para el cognitivismo y el funcionalismo, en parte como herencia de su gestación en la crítica del conductismo radical.

Por lo demás, vale la pena apuntar cómo la cuestión de la intencionalidad, sobre la que Brentano (1874) hiciera gravitar las ambiciones explicativas de la incipiente psicología, traslucía también en el fondo de las escaramuzas entre Chomsky y Skinner. No en vano, uno de las debilidades más graves que Chomsky imputa a la noción skinneriana de control estimular de la conducta verbal es su incapacidad de dar cuenta del uso de términos denotativos, como los nombres propios, en ausencia de todo contacto entre el organismo y la referencia del término tal que pueda describirse como estimulación del aquel por parte de ésta.

> Other examples of stimulus control merely add to the general mystification. Thus, a proper noun is held to be a response "under the control of a specific person or thing" (as controlling stimulus, [Skinner 1957:] 113). I have often used the words Eisenhower and Moscow, which I presume are proper nouns if anything is, but have never been stimulated by the corresponding objects. How can this fact be made compatible with this definition? Suppose that I use the name of a friend who is not present. Is this an instance of a proper noun under the control of the friend as stimulus? […] A multitude of similar questions arise immediately. It appears that the word control here is merely a misleading paraphrase for the traditional denote or refer. (Chomsky 1959: 417)

Al margen de la recurrente acusación de convertir el vocabulario conductista en una embrollada glosa de los conceptos mentalistas tradicionales, la cuestión de cómo tiene lugar el control estimular de la emisión de términos cuya referencia *de facto* no ha formado parte nunca de la estimulación que ha recibido el organismo proporcionaba a Chomsky la ocasión de enlazar con lo que, bajo esta luz, aparece como un caso particular de ese problema: el del control estimular de la emisión de términos cuya referencia no sólo no ha formado parte *de facto*, sino que *es imposible en principio* que forme parte de la estimulación recibida por el organismo, por la sencilla razón de que no existe. Junto a Eisenhower y Moscú, así pues, bien podían haber figurado la montaña dorada, Pegaso o el mayor número primo, haciendo patente de ese modo que con lo que Skinner había topado a fin de cuentas no era sino con la in-existencia intencional caracterizada por Brentano, y en cuya disciplina se emplearon ya Frege, Meinong o Russell –o, como tal vez fuera previsible, con una versión meramente empírica de esa in-existencia intencional, la ausencia fáctica en la historia estimular del sujeto.

Sea como fuera, la idea de Chomsky de que lo mental no había desaparecido del discurso conductista, sino que había quedado embozado mediante múltiples y diversos ardides, calaría pronto en el pensamiento de los pioneros del cognitivismo. En su esfuerzo por dar razón de su propia proclamación como "conductistas subjetivos", Miller, Galanter y Pribram (1960: 231, *supra*), por ejemplo, aducían sin titubear que:

> […] casi todos los conductistas han colado en su sistema algunas tretas invisibles –respuestas intervinientes, impulsos (*drives*), estímulos y qué sé yo cuántas cosas más– que

> son tan "objetivas" como lo eran en apariencia las ideas que utilizó John Locke. Todo el mundo lo hace, por la sencilla razón de que no podemos conferir ningún sentido a la conducta a menos que hagamos esto. (Miller, Galanter y Pribram 1960: 233)

Que los impulsos de Hull o las respuestas intervinientes de Tolman tuvieran tintes mentalistas no era, desde luego, revelación alguna, pero que el mismo concepto de estímulo estuviera también impregnado de mentalismo sólo podía darse por bueno una vez asimiladas las reflexiones de Chomsky.

También en virtud de argumentos de raigambre claramente chomskiana concluiría Rivière (1977) que el conductismo radical se erige sobre terrenos poco firmes. En la base del trabajo de Skinner –apuntaba Rivière incisivamente–:

> […] navega la imprecisión: aquellos conceptos a cuya precisión positiva se sacrificaron tantas cosas en el análisis experimental del comportamiento pierden su concreción en el conductismo radical. (Rivière 1977: 8)

No menos patente es la huella de Chomsky en el modo en que Pylyshyn (1984, *supra*) respalda las tesis de Brewer (1974) sobre el carácter cognitivo de los fenómenos de condicionamiento descritos en la literatura conductista. Una vez establecida la reconceptualización de la idea de reforzador como un evento informativo que permite al sujeto tomar decisiones conformes a sus creencias y deseos, y una vez anotado que el mismo efecto conductual provocado por los ensayos de condicionamiento puede alcanzarse sencillamente ofreciendo al sujeto una explicación creíble de las condiciones bajo las que recibirá distintos tipos de estimulación, Pylyshyn concede una pírrica victoria al conductista:

> I do not doubt the availability of ways of incorporating such results in the conditioning account, especially since individuation criteria for *stimulus*, *response*, or *reinforcer* are unspecified within the theory, thus allowing informal folk psychology to come to our aid in describing the situation appropriately and in smuggling knowledge of cognitive factors into the predictions. (Pylyshyn 1984: 217)

Es exactamente de ese contrabando de nociones mentalistas, amparado en la laxitud del uso que se daba en *Verbal Behavior* a los elementos básicos del aparato conceptual conductista, de lo que Chomsky (1959) acusaba a Skinner –como, en otro contexto, Chisholm (1957) y Geach (1957) habían acusado a Ryle (1949), *cf. infra*.

Entre las críticas que Chomsky, ya antes de la reseña de 1959, vertiera contra la concepción conductista del lenguaje y el problema del orden secuencial de la conducta que había advertido Lashley (1951) existía una afinidad de la que –además del propio Lashley, que llegó a plantear su argumento como la reivindicación de un modelo gramatical para la conducta motora– se percataron tempranamente Miller, Galanter y Pribram (1960), y después Baars (1986) [55], pero que no siempre ha sido

[55] La afinidad entre los argumentos de Lashley y los de Chomsky es también uno de los puntos que destaca Rivière (1991b: 141), en el contexto de un penetrante escrutinio de las concomitancias entre la

atendida. El nexo entre ambas cuestiones es patente: como apuntaban Miller, Galanter y Pribram (1960: 62-63), "[…s]eguramente, la organización de la conducta en ninguna parte podría ser tan importante como en el terreno de la comunicación […]". El propio Miller había participado en algunos intentos de describir la producción de lenguaje como conducta secuencial (Miller y Frick 1949), inspirándose en una sugerencia de Claude Shannon (1949) según la cual los procesos de Markov –*grosso modo*: cadenas estocásticas finitas en que la probabilidad del evento descrito por la variable X_{n+1} es una función de la del evento descrito por la variable X_n[56]– podrían constituir el formalismo idóneo para que la teoría matemática de la comunicación diera cuenta de la organización secuencial del mensaje. La apuesta era fuerte: "[…] si las máquinas markovianas generaran un inglés gramaticalmente correcto, resultarían igualmente adecuadas para simular todas las demás formas de conducta" (Miller, Galanter y Pribram 1960: 63). Aún a principios de la década siguiente –sólo cuatro años antes de que viera la luz su famoso estudio sobre el máximo de unidades inconexas que puede albergar la memoria a corto plazo (a saber, 7±2), que suele citarse como uno de los trabajos seminales de la psicología cognitiva (Miller 1956)–, Miller (1952) seguía intentando profundizar en las posibilidades de las cadenas de Markov para la elaboración de modelos psicológicos, descartando las basadas en matrices constantes de probabilidades transicionales si bien mostrándose esperanzado respecto al poder explicativo de las que incorporasen funciones matriciales. Pero Chomsky (1956,1957) desbarató el proyecto al demostrar que una máquina markoviana sólo podría generar la totalidad de oraciones gramaticales de un lenguaje natural si se la dotara de un número infinito de parámetros internos[57], de modo que semejante capacidad generativa –que, de hecho, bien puede exhibir un hablante competente– exigía en realidad el concurso de una gramática transformacional, cuyo poder computacional resultaba equivalente al de una máquina de Turing (es decir, una máquina que se supone capaz de resolver toda tarea que intuitivamente consideraríamos computable, *cf. infra*)[58]. Como años después resumiría Baars (1986):

---

concepción del sujeto lingüístico en Chomsky y la concepción del sujeto psicológico en las primeras formulaciones del cognitivismo.

[56] O, en las agudas palabras de Khas'minskii (2001): "[…] the 'future' and 'past' of the process are independent of each other for a known 'present'".

[57] Recientemente, sin embargo, la necesidad de integrar en las gramáticas algebraicas de inspiración chomskiana procedimientos estadísticos, a veces fundamentados en cadenas markovianas, se ha venido haciendo patente para muchos investigadores, sobre todo en el ámbito de la lingüística computacional aplicada. En el terreno teórico –se argumenta, por ejemplo, en Abney (1996)– la incorporación de modelos estadísticos sería imprescindible si aspiramos a entender cuestiones capitales como la adquisición del lenguaje, la variación lingüística, diacrónica o sincrónica, o un extenso conjunto de fenómenos que la gramática chomskiana solía desdeñar como efectos ceñidos a la ejecución de la competencia lingüística, pero irrelevantes para la descripción de dicha competencia.

[58] De hecho, Chomsky (1956) articulaba una jerarquía de cuatro clases de gramáticas, distiguiéndolas en función del tipo de lenguaje que generan, el tipo de autómata que resulta capaz de implantarlas, y el tipo de reglas que contienen. Los autómatas finitos, como los basados en cadenas markovianas, sólo serían capaces de generar lenguajes regulares; en el otro extremo de la jerarquía de Chomsky se

> Thus a Transformational Grammar, which is the *minimum* necessary gramar to account for language, is formally equivalent to the *maximally* powerful kind of theory. […] Indeed, there seems to be no viable system that can model some significant part of human behavior that is less powerful than the most powerful mathematical system. What all this means is that *the mathematical theory of automata can suggest no kind of parsimony for any system able to account for some significant part of human behavior*. All theories must be mathematically maximum theories. (Baars 1986: 177)

No puede resultar extraño, entonces, que la debacle conductista fuera percibida a menudo como una liberación de la imaginación teórica –un giro metateórico, diría Baars (1986: *passim*), que "[…] anima a los psicólogos a teorizar relativamente libres de compromisos filosóficos previos" (Baars 1986: 144), o "un nuevo permiso para hacer conjeturas" que, tal como lo veían Miller, Galanter y Pribram (1960: 67), había sido concedido a los psicólogos–, aunque esa liberación, como recordaría James J. Jenkins, viniera precedida por cierto estupor entre quienes acertaron a advertir la importancia de la debilidad de los procesos de Markov como modelos psicológicos:

> Stimulus-response models are only a subclass of Markov models. […] But English doesn't happen to be a language that can be generated by Markov models. Too bad. So that was really terribly good, it shook us all up and we all tried to work very hard on what things could be done differently. (Jenkins 1986: 246)

Parece que la devastación que el ataque de Chomsky a los modelos psicolingüístcos basados en cadenas de Markov –o, a su entender, en cualquier clase de autómatas finitos– provocaría en el seno del conductismo no era ajena a los propósitos del propio Chomsky, al menos tal como él mismo recuerda aquellos años:

> In the back of my mind was the idea that Finite Automata seemed to subsume any conceivable notion of behaviorist psychology. Given that you could show that the grammar of a language could not be represented as a Finite Automaton, it also had to be true that anything approaching the structural character of language could never *in principle* be explained by any behavioristic theory.
>
> […]
>
> So my broader interest in studying these mathematical systems was that if you could show that they're impossible in principle, then you would show that any behavioristic theory is impossible in principle. (Chomsky 1986a: 343)

Pero sin el esfuerzo de Miller los argumentos de Chomsky probablemente no habrían tenido en la comunidad psicológica el tremendo impacto que tuvieron –así lo cree también Baars (1986: 199). No en vano, Miller logró trocar la opacidad de los razonamientos formales de Chomsky en una fulgurante prueba intuitiva a la que difícilmente se podía oponer resistencia. Vale la pena, una vez más, acompañar a Miller (1986) cuando rememora aquella inflexión de su pensamiento:

---

situarían las máquinas de Turing, capaces de generar lenguajes recursivamente enumerables por medio de cualquier clase de gramática formal –es decir, de una gramática sin restricciones.

> As I thought about Chomsky's arguments, it occurred to me that if you try to learn English using purely statistical approximations to English –by using transitional probabilities between words– when you look at the size of the set of sentences 20 words long, it turns out that you have to learn an astronomical number of connections in order to generate just exactly the set of English sentences and no others. I think it works out that the average number of possible transitions following any word in a sentence is in the order of 10 –that is, at any point in a sentence there is an average of about 10 words that can follow that word. So, in sentences about 20 words long –which is not very long […]– that would lead to 10 to the 20^{th} power number of sentences. And there are less than 10 to the 10^{th} seconds in a century.
>
> So if you imagine that you have been learning one transitional probability per second since you were born, you would not have had enough time to learn more than a tiny fraction of all the sentences you can in fact produce and understand. […] At that point I was pretty well persuaded that no sort of statistical theory would ever generate what we wanted. (Miller 1986: 208)

El fruto de esta contudente reflexión, elaborado en Chomsky y Miller (1958), haría tal vez tanto como la célebre recensión de Skinner (1957) por encauzar la atención de la psicología científica hacia el trabajo de Chomsky. Acaso inadvertidamente, el razonamiento de Chomsky y Miller acabaría orientando también la crítica del conductismo lógico. Cuando Bechtel (1988: 125) pone en tela de juicio la tesis, debida a Ryle (1949), "de que los términos mentales han de hacerse equivalentes con listas potencialmente infinitas de enunciados condicionales […], puesto que tendríamos que aprender esta lista potencialmente infinita para aprender los términos mentales", la estructura del argumento es claramente deudora de la ideada por Chomsky e ilustrada por Miller. En el caso de Miller y Chomsky, el problema se contempla en un plano sintáctico: si aprender qué palabras pueden y cuáles no pueden seguir a una cadena previa pasa por aprender probabilidades transicionales brutas, el monto de probabilidades que habríamos de aprender, incluso estipulando un límite en el número de palabras que pueden formar una frase comprensible, sería desmesurado. En el caso de Bechtel, el planteamiento es de orden semántico: si la referencia de las palabras que creíamos usar para mencionar estados mentales es en realidad un conjunto de disposiciones conductuales, y dicho conjunto ha de expresarse como una lista de enunciados condicionales –o, más razonablemente, de enunciados de probabilidad– entonces parece que aprender la referencia de tales palabras equivaldría a aprender dichas listas, pero el número de enunciados que forme cada una de dichas listas carece de un límite estipulable. Las diferentes aproximaciones de Chomsky y Miller, por un lado, y de Bechtel, por otro, se justifican por la diferente naturaleza de las tesis que en cada caso se encuentran bajo asedio: Skinner (1957) está esbozando un programa de reducción de los mecanismos de aprendizaje lingüístico a los del condicionamiento operante; Ryle (1949), un programa revisionista de especificación de la semántica del lenguaje psicológico acorde al conductismo lógico.

Para los intereses de ese programa semántico del conductismo filosófico[59], parece, en efecto, considerar Rodríguez (2001b: 87) poco menos que fulminante el argumento de Bechtel, que bautiza como "el problema de la imposibilidad de aprender los términos mentales". Pues bien, conjugar ambos programas, lejos de hacer que se vigoricen mutuamente, vuelve a ambos aun más vulnerables a las objeciones de Chomsky, Miller y Bechtel. Si cada uno de los enunciados, condicionales o probabilísticos, de la lista potencialmente infinita que da, según Ryle, la referencia de un término psicológico debe ser aprendido, según Skinner, merced a procesos de condicionamiento operante –aunque, por supuesto, lo que suceda no sea estrictamente que se aprenda el enunciado como tal, sino que se fije en el organismo la tendencia a emitir el término en cuestión cuando se cumpla la condición establecida en el enunciado–, entonces, qué duda cabe, mal bastará un ensayo para hacerlo: el aprendizaje de cada enunciado será largo y costoso. Si, por otra parte, es potencialmente infinito el conjunto de pares situación – conducta descrito por tales enunciados (adviértase, por cierto, que ambos elementos del par son en el esquema skinneriano aspectos de la configuración estimular que el sujeto debe ya saber diferenciar para que el aprendizaje lingüístico sea posible), entonces mal podrá el reforzamiento de la emisión del término psicológico –o de alguno de su campo léxico o semántico, o de alguna de sus variantes morfológicas– que está siendo aprendido bajo una configuración estimular generalizarse no ya a la potencial infinidad de configuraciones equivalentes en cuanto a la descripción mentalista que soportan, sino siquiera a otras ocasiones de la misma configuración estimular situación – conducta. En este escenario, cuando –imaginemos– un niño ve a su padre silbar distraídamente mientras lo baña, y, ocasionalmente, emite alguna palabra del campo semántico de "alegría", el ocasional reforzamiento de esa conducta debe (*i*) sumarse a los posteriores episodios de reforzamiento de esa palabra, o alguna de sus variantes morfológicas, semánticas o léxicas, en configuraciones estimulares en las que la conducta de silbar de la misma u otras personas se produce en la misma situación, o en situaciones análogas en un sentido aún no descubierto por el niño –supongamos, por ejemplo, que la disposición que forma parte de la alegría es la de silbar al desarrollar actividades rutinarias–, (*ii*) sumarse también a episodios de reforzamiento de cualquiera de esas proferencias lingüísticas en un conjunto potencialmente infinito de configuraciones estimulares situación – conducta parejas a la mencionada –por ejemplo, digamos, la disposición a devolver las miradas de los demás con una sonrisa–, y, (*iii*), sumarse a los episodios de reforzamiento ocasionados por la correcta inserción del término en la cadena sintáctica, además de (*iv*) quedar modulada por los episodios de reforzamiento negativo o de castigo por emisiones inadecuadas en cualquiera de los tres ámbitos descritos.

Como se anunciaba antes, esta vertiente de los argumentos de Chomsky los enlaza, además de con la crítica del conductismo lógico, con las tempranas objeciones

[59] Respecto del cual Skinner parece haberse visto preso de cierta ambivalencia: *cf.* Skinner 1945: 415 y 1953: 268, *infra*.

de Lashley (1951, *supra*) al proyecto de Watson. Aunque en *Syntactic Structures*, donde estaba ya orquestado el ataque a los modelos lingüísticos basados en cadenas de Markov, Chomsky (1957) no citara a Lashley, la reseña de 1959 lo reconocía como precursor en varios frentes:

> […] Lashley's recognition of the problem of syntax—"a generalized pattern imposed on the specific acts as they occur" (Lashley […] 1951[…:] 119), his argument that an associative chain theory gives no understanding of the variety of integrative processes behind the production of sentences, and his suggestion that some selective mechanism must be at work in the construction of an utterance. (Bruce 1994: 99)

En un párrafo de tono marcadamente elogioso, Chomsky (1959), en efecto, tomaba nota de que ya Lahsley había dejado claro que:

> […] the composition and production of an utterance is not simply a matter of stringing together a sequence of responses under the control of outside stimulation and intraverbal association, and that the syntactic organization of an utterance is not something directly represented in any simple way in the physical structure of the utterance itself. (Chomsky 1959: 432)

Además, como bien compendia Bruce (1994), Chomsky adjudicaba sin paliativos a Lashley, citándolo, la conclusión fundamentada de que "[…] 'a consideration of the structure of the sentence and other motor sequences will show [...] that there are, behind the overtly expressed sequences, a multiplicity of integrative processes which can only be inferred from the final results of their activity ([Lashley 1951:] 509)'" (Chomsky 1959: 432), así como el reconocimiento de la dificultad que reviste la tarea de determinar "[…] 'the 'selective mechanisms' used in the actual construction of a particular utterance ([Lashley 1951:] 522)'" (Chomsky 1959: *ibid*.).

Nueve años después, Bever, Fodor y Garrett (1968) planteaban que *Syntactic Structures* constituía de hecho una generalización del argument de Lashley. Pronto se alumbraría la hipótesis de una influencia directa, que Blumenthal (1970) describe en términos de lo "profundamente impresionado" que la lectura de los trabajos de Lashley habría dejado a Chomsky. Pero Bruce (1994: 99) ha aducido, basándose en el testimonio del propio Chomsky, que su encuentro con los artículos de Lashley, que efectivamente le impresionaron, se produjo cuando había terminado ya de desarrollar sus argumentos contra los modelos basados en procesos markovianos. Esto, por otra parte, sirve a Bruce (1994) para cuestionar la preminencia en el nacimiento del cognitivismo que Gardner (1985, *supra*), y muchos otros en su estela, han otorgado al Simposio de la Fundación Hixon, en 1948, en el que Lashley pronunció la conferencia que se convertiría en el artículo de 1951, y proponer en consecuencia una relectura del papel del argumento de Lashley no como una vela que se iza en la singladura del cognitivismo, sino como una dársena en que éste recalaría posteriormente y que le proporcionaría importantísimas provisiones.

También parecen advertir Miller, Galanter y Pribram (1960) que el problema del orden secuencial de la conducta no apareció de pronto, en 1948, en el pensamiento de Lashley. Sin más elaboración ni comentario una nota al pie nos hace ver que:

> La muy conocida observación de que Karl Lashley hace en su libro *Brain Mechanisms and Intelligence* ([…] 1929), de que las ratas que han aprendido a recorrer un laberinto todavía pueden recorrerlo aun cuando Lashley les hubiera impedido, mediante operaciones quirúrgicas, utilizar la secuencia habitual de movimientos, debe querer decir que dentro de la misma estrategia general podrían sustituirse unas tácticas motoras por otras nuevas; ciertamente, la destreza del laberinto no constituye una cadena aprendida de movimientos. (Miller, Galanter y Pribram 1960: 98-99).

El dragado de las reliquias que deja en los primeros escritos de Lashley una perdurable preocupación por el problema de la sintaxis de las secuencias conductuales complejas ha sido acometido con mayor detenimiento por Bruce (1994). Hasta 1917, de hecho, se remonta la más temprana: una observación clínica relativa a un paciente "[…] who showed accurate movements of the knee in the absence of proprioceptive feedback from the knee joint" (Bruce 1994: 95). En Lashley y Ball (1929) aparecen ya los ejemplos que darían ímpetu a las conclusiones de la conferencia del Instituto de Tecnología de California, centrados en las destrezas propias de la interpretación musical. La alocución presidencial que Lashey dirigiría en 1929 a la Asociación Psicológica Americana (Lashley 1930) incidiría en los mismos asuntos; su artículo de 1937 sobre determinantes funcionales de la localización cerebral incluía una sección sobre el orden secuencial de las reacciones que puede leerse casi como un resumen de la conferencia de 1948, y su conclusión era ya entonces que:

> […] the order of the series of acts is relatively independent of the particular form of the acts which constitute it and therefore cannot be interpreted as a chain of habits. The doctrine of chain reflexes must be abandoned, and some other explanation of serial acts must be sought. (Lashley 1937: 381-382 *apud* Bruce 1994: 96)

Otro de los argumentos esgrimidos por Lashley (1951), la transferencia de secuencias entre distintos sistemas de respuesta, está ya claramente delineado en Lashley (1942), donde se muestra como ciertos rasgos de la escritura de los sujetos permanecen invariables ya se produzca ésta con la mano dominante, con la contraria, o incluso con los dientes, así como con o sin posibilidad de ver lo que se escribe, o con visión en espejo. Parece plenamente acreditado, en suma, que la insuficiencia de los esquemas básicos del conductismo para dar cuenta de las secuencias conductuales humanas complejas, lingüísticas o no, se venía larvando desde casi los tiempos del manifiesto de Watson; Chomsky (1957, 1959), dicho de otro modo, hurgaba en una herida prolongadamente abierta.

De hecho, el propio Chomsky reconocía en 1959 no sólo su deuda con Lashley (1951) –que Bruce (1994), como hemos visto, pone en tela de juicio–, sino también con

los trabajos de Verplanck (1954) y Scriven (1956). En particular, en los argumentos de Scriven había encontrado Chomsky la tesis de que "[…] in Skinner's sense of the word, stimuli are not objectively identifiable independently of the resulting behavior, nor are they manipulable" (Chomsky 1959: 435); la aportación de Verplanck a Estes (1954), en la misma línea, abundaba en "[…] the untestability of many of the so-called 'laws of behavior' and the limited scope of many of the others, and the arbitrary and obscure character of Skinner's notion of *lawful relation*" (Chomsky 1959: *ibid*.).

Divergencias y oscilaciones: las fuentes freáticas del funcionalismo

Ciertas consideraciones –quizá incluso consideraciones algo manidas– acerca del ocaso del conductismo y los albores del cognitivismo pueden alentar, así pues, la conclusión de que la psicología cognitiva no es sino la traslación de las herramientas y principios metodológicos del conductismo, con algunas añadiduras derivadas de los avances tecnológicos que propició el esfuerzo bélico de la Segunda Guerra Mundial, a una investigación de la conducta articulada desde presupuestos diametralmente opuestos –en particular, desde el presupuesto de la mediación de las representaciones mentales entre el estímulo y la respuesta. Pero una liquidación del argumento en estos términos dejaría sin atisbar siquiera otra perspectiva, seguramente de mayor calado, sobre la cuestión de la continuidad entre conductismo y cognitivismo: en qué medida el conductismo –en algunas de sus facetas al menos– era ya un funcionalismo, y en qué medida, por tanto, el cognitivismo no hizo sino esclarecer, valiéndose de herramientas conceptuales nuevas, lo que turbia o ingenuamente era ya parte de las intuiciones conductistas.

Signo de ese enraizamiento que rebasa lo metodológico pudieran ser tal vez las frecuentes referencias de los propios conductistas al carácter "funcionalista" de sus propuestas. A ojos de Watson (1913: 166), por ejemplo, el conductismo aparece como "el único funcionalismo consistente y lógico". Por supuesto, lo que Watson designa como "funcionalismo" es la forma de entender lo mental que él mismo había aprendido de Angell durante sus estudios en la Universidad de Chicago, en los primeros años del nuevo siglo, cuando aún postulaba la existencia de "impresiones intraorgánicas" (Watson 1907: 84-85 *apud* Morris y Todd 1999: 27) para explicar la conducta en el laberinto de ratas a las que había infligido severas lesiones perceptivas –un aprendizaje más lento, pero con el mismo rendimiento final. De modo que lo que estaba presente en el texto que –siguiendo la intuición historiográfica de Boring (1950) y su cuidadosa reivindicación por parte de Burnham (1968)– hemos dado en reconocer como acta fundacional del movimiento conductista –a saber: el manifiesto de febrero de 1913–, era cierta constatación del enraizamiento de la nueva visión de la psicología allí enarbolada en la tradición funcionalista proveniente de William James. Prosigue Watson, de hecho, argumentando que sólo el conductismo puede enderezar el rumbo de esa tradición entre "la Escila del paralelismo y la Caribdis de la interacción […], reliquias […] largo tiempo veneradas

por la especulación filosófica" (Watson 1913: 166) que él propone sencillamente ignorar.

Ahora bien, los trazos de continuidad entre el propio funcionalismo cognitivista y la concepción de la mente de James o Angell son suficientemente profusos, sobre todo si se da por buena la necesidad de una reinterpretación teleológica de la teoría funcionalista de la semántica de los estados mentales (*cf.* Bechtel 1988: 185, *infra*), como para autorizar una tentativa indagación de cuánto de lo que Watson toma de Angell pervive luego, transformado, en las formulaciones pioneras del proyecto cognitivista. Rastrear esos rescoldos cobra además cierta motivación simbólica si reparamos en la condición de primer crítico del conductismo que reviste a Angell. El mismo volumen de *The Psychological Review* que publicara el manifiesto de Watson incluía también, en el número de julio, la advertencia, proveniente de una conferencia dictada por Angell en el congreso anual de la Asociación Psicológica Americana del 31 de diciembre del año anterior, de que la condena de la consciencia a la que nos espoleaban materialistas, comparatistas y objetivistas no podía sino abocarnos a obviar aspectos cruciales de lo que la psicología legítimamente anhelaba averiguar. Entre las fuerzas que nos aguijaban hacia la condena de la consciencia se contaba, a su entender, "[…t]he revolt against the domineering claims of introspection as the alpha and omega of psychological method" (Angell 1913: 257), una revuelta que –*nota bene*: apenas dos meses antes de que viera la luz el manifiesto de Watson– no representaba aún "ningún programa formalmente reconocido" (Angell 1913: 257). El propio Angell –que interpretaba en esa tendencia innominada signos de una "base más duradera y sustancial" (Angell 1913: 257)– se atrevería, como de pasada, a bautizarla: "Let us imagine our psychologist –or our behaviorist, as we may shortly call him– starting to define his field […]" (Angell 1913: 261). Pero el creciente desuso de la noción de consciencia en la explicación psicológica –según había presagiado Angell ya en 1910– no entrañaría "[…] the disappearence of the phenomena we call conscious, but simply the shift of psychological interest toward those phases of them for which some term like behavior affords a more useful clue" (Angell 1913: 255). Las ocasionales, inseguras derivas de Watson hacia los piélagos de una interpretación metafísica del conductismo, donde se apiñaría con las conclusiones del conductismo lógico, pueden entenderse, entonces, como momentos en los que desoyó esta observación de su antiguo profesor.

Entre los aspectos del pensamiento de Watson que con mayor claridad resuenan en los planteamientos cognitivistas cabe contar su decidida defensa de la autonomía de la psicología respecto de la fisiología, por la que Watson se había enfrentado a Jacques Loeb, otro de sus maestros en la Universidad de Chicago. La afirmación de que las unidades en que habría de desplegarse el análisis de la conducta eran distintas de las de la investigación neurofisiológica –así, exenta– bien podría ser respaldada por muchos cognitivistas –*tendría* que serlo, al menos, por quienes no estuvieran inmersos en un cuestionamiento de las intuiciones que imprimieron su fuerza inicial al giro cognitivo de la investigación psicológica. Con

parecido espíritu, por otra parte, Skinner se anticipa al papel decisivo que en el desarrollo de la concepción funcionalista de lo mental tendría la distinción entre, por una parte, un materialismo comprometido sólo con la tesis de que todo estado de la mente sea de hecho un estado del cerebro y, por otra, un materialismo mucho más ambicioso, y más afín a las ínfulas de universalidad del vocabulario fisicalista, según el cual toda clase de estados mentales sería idéntica a una clase de estados cerebrales (*cf. infra*). No en vano, ya en *Ciencia y Conducta* Humana Skinner se percata con toda claridad de que:

> […n]o es correcto decir que el refuerzo operante 'fortalece la respuesta que la precede'; la respuesta se ha producido ya y no puede cambiarse. Lo que cambia es la probabilidad futura de respuestas de la misma clase. Una operante es una clase de conducta. (Skinner 1953: 104)

Años atrás, en su estudio de los conceptos básicos del condicionamiento pavloviano, el propio Skinner (1935) había dejado anotado que:

> […i]n a reflex preparation the observed correlation is never between all properties of both stimulus and response. Some properties are irrelevant. The relevant properties are accordingly taken to define classes and the reflex is regarded as a correlation of generic terms. (Skinner 1935: 476-477)

Sin titubeos, Skinner trazaba inmediatamente una conclusión que sólo se aparta de los planteamientos funcionalistas –aparte, claro, de la distancia que impone la aversión de Skinner a aplicar su análisis al concepto de estado mental– en el peculiar modismo –"nivel de restricción"– elegido para referirse a lo que un cognitivista llamaría sin ambages "nivel de abstracción":

> A reflex is accordingly defined as a correlation of a stimulus and a response at a level of restriction marked by the orderliness of changes in the correlation. (Skinner 1935: 477)

Cuando abordasen la cuestión de la unidad de análisis apropiada para la investigación psicológica, Miller, Galanter y Pribram (1960: 31-32) encomiarían el enfoque adoptado por Skinner para la definición del concepto de reflejo como "[…] el resultado del coherente intento skinneriano de definir una unidad conductual en términos de la conducta misma, en lugar de hacerlo por referencia a conceptos extraídos de alguna otra rama de la ciencia" –pese a que, por lo demás, trataran de deshacerse de una adhesión al propio concepto de reflejo que consideraban desmesurada (*cf.* Miller, Galanter y Pribram 1960: 16-17, *infra*). Desde la perspectiva que da el tiempo, mucho más rotundo se mostraría Mandler (2002: 341) en su valoración de la continuidad entre ciertos aspectos del trabajo de Skinner y la concepción funcionalista de lo mental: sencillamente, Skinner había abandonado la determinación de Watson de definir "[…] organism response in terms of its physical parameters", y había comenzado decididamente a usar "[…] functionalist definitions

of stimuli and responses as eliciting/discriminative conditions and operant behavior"[60]. Veremos, no obstante, que muchos pensadores de la órbita funcionalista han permanecido más apegados a Watson que a Skinner en este punto, restringiendo la definición funcional al ámbito de los estados mentales e insistiendo en una definición física de estímulos y respuestas –*cf*., por ejemplo, Block (1978: 64) o Armstrong (1968: 84), *infra*.

Más a primera vista: cuando Skinner se niega a identificar determinada variable independiente de sus paradigmas experimentales con el hambre *entendida como estado mental*, subjetivo, parece acampar en las antípodas del cognitivismo. Pero si atendemos a que el mismo Skinner –como recuerda, por ejemplo, Boring (1950: 672)– rechazaría también identificar dicha variable con el hambre *entendida como estado fisiológico* o físico del organismo, lo encontraremos de pronto mucho más cerca de los planteamientos funcionalistas. Del profesor William J. Crozier, un discípulo rebelde de Jacques Loeb en cuyo laboratorio de fisiología, en Harvard, se había desempeñado como ayudante, Skinner había aprendido –según su propio testimonio– "[…] el desprecio […] hacia la 'fisiología de los órganos' de la escuela de medicina" (Skinner 1978: 114). Pero –qué duda cabe– aquel aspecto de su posición ha sido mucho más recalcado que éste último –que consigna impecablemente, entre otros, Gondra (1992: 39).

Igualmente revelador, si bien en sentido opuesto, resulta rememorar los términos bajo los que Albert P. Weiss (1925/1929) otorgaba a un psicólogo el entonces muy apreciado título de conductista, que podrían trasladarse *a contrario*, casi punto por punto, a una descripción del sustrato funcionalista del cognitivismo:

> Behaviorism at present is merely a convenient term which more or less definitely separates those psychologists who believe that the so-called mental states cannot be classified as physical states, from those psychologists who believe that they can. (Weiss 1925/1929: 5).

Al menos desde la óptica antirreduccionista que la huella de Putnam y Fodor ha hecho tan corriente en la interpretación epistemológica del cognitivismo, la demarcación trazada por Weiss sirve no sólo para expulsar del ámbito del conductismo a los psicólogos que entonces se mantenían ligados a la tradición wundtiana, sino también a los psicólogos cognitivos que acabarían con la hegemonía conductista. Es, en efecto, uno de los principios fundamentales de esa interpretación

[60] Que Chomsky (1959) errara al blandir "[…] sentimientos jacobinos" contra Skinner (1957) porque la postura de Skinner era ya netamente funcionalista, como sugiere Mandler (2002: 343), exige, sin embargo, forzar en exceso la lectura de Skinner: una cosa es que las definiciones de estímulos y respuestas construidas por Skinner fueran definiciones funcionales, y otra bien distinta que Skinner aceptara el uso de la definición funcional para introducir en la explicación psicológica mecanismos de teorización sobre estados mentales. Lo que denunció Chomsky es, precisamente, que Skinner (1957) incorporaba veladamente teorización sobre estados mentales a sus explicaciones, aun cuando rechazase con vehemencia hacerlo. Bien es cierto, eso sí, que el tono de la reseña era, como apunta Rivière (1991b: 142), algo desabrido.

antirreduccionista de la psicología cognitiva que los estados mentales –los "así llamados estados mentales", si se quiere preservar el matiz peyorativo de Weiss– no pueden ser clasificados como estados físicos.

No obstante, basta detenerse siquiera someramente en el análisis del pensamiento de Weiss para apreciar como las raíces del cognitivismo ahondan incluso en planteamientos conductistas con los que parecen a primera vista irrenconciliables. El compromiso de Weiss con el reduccionismo psicofísico era mucho más estricto que el de Watson o Skinner. La preocupación por la traducción del vocabulario psicológico subjetivo a términos objetivos había calado en Weiss –como bien ha mostrado Wozniak (1994)– casi desde el día en que su carrera académica se orientó un poco azarosamente hacia la psicología, debido a un encuentro con el profesor Max F. Meyer, un discípulo de Stumpf y –*rara avis*– Ebbinghaus, a la sazón en la Universidad de Missouri. La obra de mayor envergadura de Meyer, *The Fundamental Laws of Human Behavior* (1911), excluía tajantemente de la explicación psicológica el uso de conceptos subjetivos –*i.e.*, referidos a estados mentales– *salvo* que se tomaran como abreviaturas de procesos nerviosos complejos. Con ello, Meyer apuntaba una actitud ante la investigación psicofisiológica muy diferente de la renuencia con que habrían de afrontarla Watson o Skinner, y también una concepción del papel del vocabulario teórico en ciencia algo más afín a lo que habría de ser el positivismo lógico que –como sucede en Watson o Skinner– a un positivismo ingenuo de raigambre comtiana o baconiana. Cuando Weiss empezó a desempeñarse como ayudante de laboratorio de Meyer, el profesor estaba trabajando precisamente en la redacción del tratado que la historiografía conductista saludaría como "la primera explicación conductista completa de la acción humana" (Pillsbury 1929: 290 *apud* Wozniak 1994: *ix*) –el propio Pillsbury, de hecho, había sido seguramente el primero en definir la psicología como "el estudio de la conducta", en las líneas iniciales de su *Essentials of Psychology* (1911).

Pero además de sus empedernidas convicciones objetivistas y reduccionistas, Weiss aprendió de Meyer a poner el acento en la dimensión social de la conducta –a juicio de Meyer, lo único que sustentaba una aproximación psicológica al fenómeno en lugar de una estrictamente fisiológica (*cf.*, por ejemplo, Meyer 1921: 405). En la medida en que constatamos sin atisbos de controversia que existen "[…] two criteria with respect to which human movements can be classified: (a) as neuromuscular effects of *preceding* movements, (b) as neuromuscular causes of *subsequent* movements" (Weiss 1925: 55-56) –clasificación biofísica y biosocial, las denomina Weiss–, la psicología estaría llamada a cubrir el hiato entre las ciencias naturales y las ciencias sociales. Ahora bien, su riguroso reduccionismo no impide a Weiss darse perfecta cuenta de que respuestas pertenecientes a la misma clase biofísica pueden pertenecer a distinta clase biosocial, y viceversa, así como de que, en consecuencia, ninguno de los dos sistemas de clasificación es prescindible en favor del otro. De hecho, la concepción de Weiss de la relación entre lo biosocial y lo biofísico transita a veces muy cerca de la concepción funcionalista de la relación entre lo psicológico y lo neurológico: leemos, por ejemplo, que "[…] the individual is classified not on the

basis of physical or physiological properties but on the basis of his co-operative status in the social organization of which he is a unit" (Weiss 1925: 142), o vemos a Weiss rechazar sin titubeo la tesis de que "[…] neurological insight alone will enable us to determine what the *stimulating effects* of a given neuro-muscular configuration will be *upon other individuals.*" (Weiss 1925: 81-82). Desde luego, no esquiva al lúcido análisis de Wozniak (1994: *xv*) el hecho de que Weiss ha introducido en su fundamentación teórica de la psicología conductista la noción de niveles de discurso teórico, abriendo así la posibilidad de que unos niveles de discurso resulten no ser reducibles a otros. Además, Weiss se percata de que nada de esto le obliga a admitir que coexistan dos realidades, física y social, en los fenómenos así descritos, ni siquiera dos *aspectos* de una misma realidad: se trata ni más ni menos que de dos taxonomías. Precavido del riesgo de ser tildado de dualista, apunta:

> I wish to direct special attention to the fact that biophysical and biosocial do *not* refer to two *aspects* of the same thing, say the type of sensorimotor organization. The classifications which include responses that are biosocially similar need not show any biophysical similarity (Weiss 1924: 42-44)

Lo que Weiss ha advertido con claridad es que el reconocimiento de la distancia entre la clasificación biofísica y la biosocial no debilita su monismo. Ésta es a todas luces, trasladada a la relación entre lo psicológico y lo fisiológico, la misma intuición que impulsaría en algunos círculos el abandono del fisicalismo –de la identidad psicofísica, si por tal se entiende que cada tipo de estado mental sea en realidad un tipo de estado cerebral– y el desarrollo de una alternativa funcionalista –basada en un materialismo más parco, que asume sólo que cada estado mental particular es de hecho un estado cerebral particular– como fundamento de la autonomía de la psicología cognitiva respecto de las neurociencias (*cf. infra*). Lo que Weiss acaso no llegara a advertir con claridad es que el reconocimiento de la distancia entre la clasificación biofísica y la biosocial sí podría debilitar su reduccionismo, en la medida al menos en que la interpretación antirreduccionista del funcionalismo se muestre finalmente viable. Un giro semejante habíamos advertido en el caso de Skinner(1935: 477, supra), cuando la definición de "reflejo" queda sujeta a un concepto de "nivel de restricción" que se anticipa a la insistencia funcionalista en diferenciar niveles de abstracción seguramente con menor perspicuidad que la alcanzada por Weiss en su aproximación a la idea de niveles de discurso teórico. Lo que todo esto parece apuntalar es que existía en el seno del movimiento conductista una tensión constante, y presente en ámbitos muy diversos, hacia los planteamientos que a la larga conformarían la ortodoxia cognitivista.

Parece claro, en fin, que en el seno del conductismo cabe el hallazgo de concepciones irremediablemente alejadas de las del cognitivismo –eso no es noticia–, pero también de otras cuya afinidad con los planteamientos que alumbrarían los nuevos modelos cognitivos es rotundamente destacable, y que a veces anidan incluso en los territorios más hostiles, más declaradamente comprometidos con la reducción

psicofisiológica o con la eliminación del vocabulario subjetivo. Pero acentuar en exceso las facetas más rupturistas del proceso es –qué duda cabe– una propensión que se registra con frecuencia en los procesos de cambio histórico, en ciencia como en cualquier otro ámbito de actividad humana.

En el terreno, por otra parte, de la postulación de procesos internos en la determinación de la conducta, sea cual sea el dictamen sobre la autonomía explicativa de la psicología, tanto Watson como Skinner se mostraron en todo momento mucho más taxativos, aunque sus prácticas de descripción de estímulos y respuestas (*cf. infra*) desautorizaran con frecuencia sus pronunciamientos. Por el contrario, neoconductistas tan destacados como Tolman (1959: 98) o Guthrie (1959: 165) reconocerían abiertamente, en Koch (1959b), no sólo el carácter cognitivo de sus constructos teóricos sino, además, que es eso "[…] lo que confiere significado psicológico […] a los estímulos y las respuestas" (Yela 1980/1996: 175). Pero entonces –se apresuraba a anotar el propio Koch (1959d: 769)– la distinción entre el lenguaje objetivo de estímulos y respuestas y el que los conductistas habían tachado de subjetivo pierde todo su valor epistemológico. El cognitivismo no podía ya ser impugnado por razón de presuntas flaquezas epistemológicas.

De cualquier modo, en tanto que puntas de lanza de una concepción de lo mental cuyas posiciones metafísicas registraron significativas oscilaciones, los planteamientos de Watson, como después los de Skinner, son particularmente confusos. Tanto uno como otro, por ejemplo, cometieron el error de alinearse por motivos tácticos con la doctrina epifenomenista, según parece sin advertir que postular la existencia de un conjunto de fenómenos psicológicos que se diferencian de los fenómenos físicos por su menguada eficacia causal es en realidad tanto como comprometerse, *velis nolis*, con una posición dualista. Con tono displicente, Watson (1913) reconocía que:

> I may have to grant a few sporadic cases of imagery to him who will not be otherwise convinced, but I will insist that the images of such a one are sporadic, and as unnecessary to his well-being and well-thinking as a few hairs more or less in his head (Watson 1913: 423)

Pero esporádicas u ordinarias, superfluas o decisivas, unas pocas imágenes mentales son ni más ni menos que eso: imágenes mentales. Para colmo de males, la interpretación netamente reduccionista, o eliminacionista, del pensamiento de Watson sería blanco de acres arremetidas por parte de los defensores estrictos de una concepción epifenomenista de lo mental:

> […] el error de Watson fue que, para demostrar que no hay mentes que interactúan, lo que es verdad, creyó necesario afirmar que *no hay mentes*, lo que no sólo es falso, sino estúpido. (Bergmann 1956: 266 *apud* Yela 1980/1996: 169)

Del mismo error se vería preso Skinner (1953) al argumentar que la psicología debía renunciar a mencionar en sus explicaciones hechos subjetivos, mentales, *porque*, entre

otros motivos, la naturaleza estrictamente privada de tales hechos impediría establecer sus relaciones funcionales con estímulos y respuestas, en el supuesto –a su juicio, ininteligible– de que las tuviera. Aun dando por bueno el argumento –lo que sería tanto como desabastecernos de los mecanismos de introducción de términos teóricos de los que el positivismo lógico nos había dotado incluso para fenómenos no observados ni siquiera privadamente y regresar a una concepción comtiana de la explicación científica– resulta claro que Skinner asume tácitamente la existencia de los presuntos hechos subjetivos. Pero aunque lo hiciera sólo por razones retóricas, a fin de poder negarles a renglón seguido toda eficacia causal, un hecho subjetivo no deja de ser un hecho subjetivo por muy desprovisto de eficacia causal que quede, de modo que los razonamientos de Skinner menoscaban toscamente, una vez más, su propia posición materialista. La estructura argumentativa con la que Skinner trabaja en este terreno, y que reaparece de forma más o menos nítida en varios de sus escritos de crítica del cognivismo (Skinner 1977, 1985, 1989, 1990), ha quedado limpiamente diseccionada en Fernández Trespalacios (1992). El punto de partida es denegar taxativamente la existencia de realidades mentales –estructuras, procesos, estados, etc.–; después, se concede por mor del argumento su existencia, aunque sólo para negar su relevancia explicativa –o su eficacia causal–; por último, se concede, nuevamente por mor del argumento, la relevancia explicativa de la postulación de realidades mentales, para cuestionar inmediatamente la utilidad de las explicaciones así articuladas, apelando a la dificultad de detectar y describir los fenómenos mentales cuya existencia se había comenzado por rechazar. Así vista, la acusación que pesa contra Skinner gravita sobre dos cuestiones: Skinner concede más de lo *cree* conceder, y desde luego más de lo que *puede* conceder *sin perjuicio* para sí.

En *Sobre el conductismo* (Skinner 1974) el carácter privado de los fenómenos psicológicos pierde fuerza en tanto que razón para obviarlos en la explicación de la conducta. Es, más bien, el hecho de que tales fenómenos no constituyen causas de la conducta lo que nos obliga a expulsarlos de su explicación, y nos obligarían a hacerlo –según parece– incluso si la introspección nos los mostrara con toda fidelidad. Con esta maniobra, Skinner se adentra más aun en el territorio del epifenomenismo. Así, por ejemplo –como señalaba Rivière (1977: 2)– los sentimientos son para Skinner (1974: 47) "[…] simplemente productos colaterales de las condiciones responsables del comportamiento", ya que "[…] el cambio en los sentimientos y el cambio en el comportamiento tienen una causa común" (Skinner 1974: 60) –una tesis, en fin, nítidamente epifenomenista. Bajo esta luz resulta inteligible la deriva de las preocupaciones de Skinner hacia la desarticulación y reconstrucción de la idea de libertad, deriva que había tenido su culmen en Skinner (1971): la clave de arco de su proyecto ha pasado a reposar sobre la tesis de que "[…] la persona no es un agente generador, es un *locus*, un punto en el cual confluyen muchas condiciones genéticas y ambientales en un efecto común" (Skinner 1974: 152). Pero resulta también censurable que las asfixiantes restricciones metodológicas urgidas por Skinner acaben –como sagazmente anotara Rivière– dictándose:

> […] en aras de un causalismo extremo que no está justificado ni por el modesto concepto de "relación funcional" de que parte Skinner ni por los resultados que en función de probabilidad se obtienen en el laboratorio de análisis experimental del comportamiento. (Rivière 1977: 7)

Parece que Watson osciló asimismo a lo largo de toda su carrera académica –tal como han admitido Morris y Todd (1999: 36)– entre propugnar la inexistencia y la irrelevancia de lo mental. Vacilaba al respecto en el manifiesto de 1913, y persistía en el titubeo incluso en su obra de madurez, donde "[…] no respondió directamente a la pregunta de si [el conductismo] era un mero 'enfoque metodológico para el estudio de problemas psicológicos o un sistema real de psicología'" –así lo recordaba Gondra (1992: 31), evocando palabras del propio Watson (1925: 16). Skinner, por su parte, tan pronto aseguraba haber proporcionado "[…] un cumplido análisis operacional de los conceptos mentalistas tradicionales" (Skinner 1945: 415), con la misma confianza implícita en la psicología ordinaria que habían mostrado Tolman (1932) o Hull (1943), como se jactaba, con un tono claramente eliminacionista, de que al conductismo le bastaba con desestimar tales conceptos. Al mirarla de frente, una conclusión estrictamente plegada a los principios del conductismo lógico, como la identidad entre la sensación y las disposiciones comportamentales que solemos pensar que sustenta, resultaba para Skinner claramente rechazable, convencido como estaba de que:

> […] hemos de evitar llegar a la dudosa conclusión de que, en la medida en que concierne a la ciencia, el informe verbal, o cualquier otra respuesta discriminativa, *es* la sensación. (Skinner 1953: 268)

Pero tal vez cuando insistía en extender su proyecto de análisis operacional a todas las esferas de la conducta humana Skinner pasaba por alto tal convicción, aunque fuera sólo momentáneamente. Así, de hecho, lo denunciaría Chomsky en la demoledora recensión de *Verbal Behavior*: mientras que en *The Behavior of Organisms*, Skinner (1938) había asumido, más cuidadosamente, que "[…] the terms of casual description in the popular vocabulary are not validly descriptive until the defining properties of stimulus and response are specified, the correlation is demonstrated experimentally, and the dynamic changes in it are shown to be lawful" (Chomsky 1959: 435), el núcleo del proyecto de *Verbal Behavior* era precisamente el trasvase del vocabulario empleado en la descripción experimental de fenómenos de condicionamiento a la explicación de la conducta verbal humana en el ambiente en que naturalmente se produce, desligando de esa forma dicho vocabulario de las restricciones metodológicas propias de su ámbito originario e incardinándolo en los patrones, más laxos, de la "descripción casual". Tenía, pues, toda la razón Skinner (1938: 41-42) al aseverar que "[…] in describing a child as hiding from a dog, 'it will not be enough to dignify the popular vocabulary by appealing to essential properties of *dogness* or *hidingness* and to suppose them intuitively known'" (Chomsky 1959: 435); por desgracia, esa fallida dignificación de la explicación psicológica de sentido

común, de acuerdo con el análisis de Chomsky, es ubicua en Skinner (1957). *Verbi gratia*: "[…] our belief in what someone tells us is similarly a function of, or identical with, our tendency to act upon the verbal stimuli which he provides" (Skinner 1957: 160 *apud* Chomsky 1959: 419). El intento de traducir el vocabulario mentalista de la psicología natural a la jerga conductista es aquí tan patente como puede llegar a ser en Skinner, empañado siempre por su ambivalencia ante las conclusiones del conductismo lógico –ante su método, el análisis lógico del lenguaje ordinario, Skinner se mostraría tan reacio como ante el del positivismo lógico, a saber, el análisis lógico del lenguaje científico. Así, que la creencia sea una disposición conductual o una función de dicha disposición parecen constituir para Skinner enunciados equivalentes, lo que claramente se perfila como una tentativa de eludir la cuestión.

Acaso con vocación más ecuánime, el propio Skinner (1974: 19) daba por bien entendido que "[…] algunos [términos mentalistas] pueden traducirse por [términos referidos a] comportamiento, otros se pueden descartar como innecesarios o inútiles"; dónde trazar la línea, naturalmente, habría de dirimirlo la investigación experimental. Sin embargo, incluso entonces –casi parece que para avivar la confusión– Skinner rechaza que su análisis quede caracterizado como reduccionista, ni siquiera para los términos traducibles: el conductismo –se nos dice– "[…n]o reduce los sentimientos a estados corporales; simplemente dice que lo que se siente son, y siempre han sido, estados corporales" (Skinner 1974: 217). No es fácil ver, desde luego, qué otra cosa entiende Skinner que sería reducir los sentimientos a estados corporales –si, como hace ver Rivière (1977: 5) la noción de reduccionismo ha adquirido "un matiz peyorativo en psicología", impúgnese el matiz, pero no se busque refugio en él para disimular los propios compromisos ontológicos. De todos modos, la maniobra es pasajera: a renglón seguido nos advierte Skinner de que el conductismo "[…n]o *reduce* a comportamiento los procesos de pensamiento, simplemente analiza el comportamiento que antes se explicaba con la invención de procesos de pensamiento" (Skinner 1974: *ibid.*), abandonando de nuevo ese renuente reduccionismo para abrir paso a un indiscriminado eliminacionismo respecto del pensamiento.

Discrepancias semejantes se daban entre destacados adalides del proyecto conductista. Como recordaría Boring (1950), ya Holt (1914) –como Tolman, o Hull– había dejado "[…] claro que el conductismo no exorciza a la conciencia sino que la absorbe" (Boring 1950: 677), y Weiss (1925/1929), en cambio, que:

> El conductismo afirma que posee una explicación *más* completa y *más* científica de la totalidad de los logros humanos *sin* la concepción de la conciencia que la psicología tradicional *recurriendo* a ella. Los factores que la psicología tradicional vagamente clasifica como conscientes o elementos mentales, simplemente *desaparecen* sin dejar ningún reducto en los componentes biológicos y sociales del análisis conductista. (Weiss 1925/1929 *apud* Boring 1950: 670)

Aunque la divergencia entre Holt y Weiss es patente –bien podría servir, de hecho, para ilustrar el contraste entre una concepción reduccionista y una eliminacionista de la relación entre la psicología ordinaria y la explicación científica–, Boring no sólo no repararía en ella, sino que destacaría la capacidad del enfoque de Weiss de trascender ese contraste. Que Watson habría carecido de tal capacidad quedó más rotundamente proclamado por Boring que en qué consistía exactamente lo que Watson no supo hacer –puesto que no se trataba, al parecer, de tomar partido. La metáfora digestiva elegida por Boring dista de ser clarificadora:

> […L]os conductistas sofisticados […c]onservan la conciencia haciéndola objetiva. Abandonan la terminología mentalista y trabajan sólo con datos objetivos de entidades sociales o físicas *o*, como Tolman, introducen variables intervinientes que se reducen a datos objetivos cuando se consideran las operaciones de sus observaciones […]. La ingestión conduce a la absorción. (Boring 1950: 671, énfasis añadido)

Pero lo que se discute es precisamente que los dos términos de la disyunción planteada por Boring sean sinónimos –como él parece dar por sentado–, o más bien antónimos, pues, *in nuce*, *abandonamos* –eliminamos– los términos teóricos de las teorías irrecuperablemente erróneas y en cambio *reducimos* a nuestras nuevas teorías los de las teorías que resultan al menos parcialmente verdaderas. Ni siquiera, por otra parte, puede asumirse que la posición de Weiss fuese de hecho el eliminacionismo estricto que Boring parece entresacar de sus palabras –aunque sin registrarlo. Antes al contrario, parece más acertado interpretar –en la línea de Wozniak (1994)– que el conductismo de Weiss reconocía "[…] the existence of conditions for which mental terms are employed but strives to represent these conditions in terms of their objective equivalents" (Wozniak 1994: *xiii*); es decir, que Weiss –como Holt, y como luego Tolman o Hull, pero sobre todo como su maestro Meyer– se pronunciaba tajantemente y sin fisuras en favor de la idea de reducción operacional. Incluso la identidad entre sensaciones y respuestas discriminativas, que Skinner (1953: 268, *supra*) había rechazado tajantemente, encontraría un valedor en Ribes (1990), quien –como bien señala Campos-Roldán (1999), bajo el ostensible influjo de Ryle– sostiene que:

> Todo proceso lingüístico de referencia a eventos «privados» o «subjetivos» […] [es] el evento privado de referencia [...]. Los eventos privados son formas particulares de contenidos descriptivos que se emiten en relaciones públicas, y nunca acontecimientos […] que determinen dichas descripciones como correspondencias evento-significado (Ribes 1990: 60-61).

En otros momentos, por lo demás, Skinner adoptaría un talante más conciliador, apoyándose en la distinción entre discurso ordinario y jerga para respaldar el uso coloquial de conceptos mentalistas –incluso en la práctica de la psicología clínica– sin atender a veleidades reduccionistas ni eliminacionistas:

> We need a language of feelings and states of mind in our daily lives. It is the language of literature and most of philosophy. Clinical psychologists use it to learn many things about the histories of their clients that they could not discover in any other way. There are two languages in every field of knowledge, and it would be foolish to insist that the technical version always be used. But it must be used in *science* and specially in a science of behavior. (Skinner 1985: 300)

No muy distantes son los predios que atraviesa Watson cuando, ya en 1913, concede que la consciencia, a la que nos exhorta a desterrar de la investigación psicológica es, en el sentido más cotidiano, precisamente lo que hace posible tanto dicha investigación como cualquier otra investigación científica, o –cabe añadir– de cualquier otra índole:

> Psychology, as the behaviorist views it, is a purely objective, experimental branch of natural science which needs introspection as little as do the sciences of chemistry and physics. […] We might call this the return to a non-reflective and naive use of consciousness. In this sense consciousness may be said to be the instrument or tool with which all scientists work. (Watson 1913: 176)

En realidad, los titubeos de Watson, como después los de Skinner, en cuanto a si las razones por las que debemos desertar del estudio de la consciencia son de índole sólo metodológica o también ontológica –titubeos que, según señalan Morris y Todd (1999: 38), aún persistían en la que sería la expresión más madura de su pensamiento, la segunda edición de *Behaviorism* que publicó en 1930– minaron incluso en sus días de esplendor la credibilidad del conductismo. Como han observado también Morris y Todd:

> In the final analysis, Watson was probably a metaphysical behaviorist, but his position is difficult to discern because most of his arguments were methodological, not metaphysical […].
>
> As a result, although Watson was praised for his commitment to objectivity, his incompletely developed rejection of consciousness was evidence for the necessity of retaining it. He seemed to have underestimated how deeply pervasive the concept of consciousness was in psychology. (Morris y Todd 1999: 36-37)

Irónicamente, es de retener bajo el abrigo del concepto de mente prejuicios metafísicos provenientes de la idea de alma de lo que Watson había acusado a los psicólogos que trabajaban en la estela de Wundt:

> When the psychologist threw away the soul he compromised with the conscience by setting up a "mind" which was to remain always hidden and difficult to access. The transfer from the periphery to cortex has been the inventive for driving psychology into vain and fruitless searches of the unknown and unknowable. (Watson 1913: 424)

Pero el sino de Watson o Skinner, como el de Ryle, habría sido –si todo esto es así– el de retener bajo el abrigo de los conceptos de estímulo y respuesta ciertos prejuicios

psicológicos provenientes de la idea de mente que vanamente se habrían afanado en socavar. Si la caída en desgracia del conductismo, como argumentaba Mackenzie (1977), era intrínsecamente inevitable, habría de ser cierto también que las simientes que la engendraron fijaban de antemano el curso de la concepción de lo mental que estaría llamada a reemplazar al proyecto de Watson –o acaso, como sostiene Reed (1997), lo torcían de antemano. Pero, una vez más, que los hechos históricos –ni siquiera en la historia de la ciencia– se ahormen por fuerza a necesidades lógicas es sumamente inverosímil: hace falta mirar para ver.

El dolor y la fragilidad: la naturaleza de las disposiciones en el conductismo lógico

Acaso por ser sus compromisos ontológicos algo más nítidos que los de Watson o Skinner, la continuidad entre la concepción de lo mental del conductismo y el cognitivismo puede aquilatarse con mayor exactitud atendiendo al diálogo que el incipiente funcionalismo entablara con la figura de Gilbert Ryle. Un apretado compendio de las fisuras que a ojos de los pioneros del cognitivismo hacían indefendible el conductismo lógico puede hallarse en Rabossi (1995): así, las transcripciones del vocabulario psicológico a oraciones relativas a conductas y circunstancias observables, que Ryle (1949) anunciaba como rédito de la ruptura con el cartesianismo, "[…] resultan demasiado generales y, en consecuencia, dudosamente elucidatorias, o bien implican un número no acotable de circunstancias específicas", pero es que además –apunta Rabossi– "[…] los análisis de tipo conductista sólo son posibles si se admiten ciertos supuestos acerca de los estados mentales de los agentes" toda vez que "[…] dos agentes pueden diferir en sus estados psicológicos pese a la similitud de sus respuestas conductuales" –como dramática e imaginativamente había demostrado Putnam (1963a, *infra*) a la vez que ayudaba a afianzar la intuición de que "[…] hay estados mentales no acompañados por tales respuestas"–, y, por último, "[…] no resulta admisible negar la relevancia de la ciencia respecto del problema mente-cuerpo" (Rabossi 1995: 21), el cual, a juicio de Ryle, se disolvía en un terreno estrictamente lógico. Ciertamente, como aventura Rabossi al señalar los senderos por los que habían de dirigirse las indagaciones acerca de la naturaleza de lo mental toda vez que los penúltimos baluartes conductistas quedaran abandonados *por estas razones*, el devenir de la tesis de identidad psicofísica, y a la larga del funcionalismo, estaba ya en buena medida prefigurado en las carencias que se reprochaban al conductismo lógico:

> Si se rechaza el conductismo en base a *ese tipo de argumentos*, se debe estar dispuesto a sostener la relevancia de los estados internos de los agentes y sus conexiones causales, a dudar del interés teórico de los planteos filosóficos que se centran en el análisis del significado de los términos y oraciones mentales, a explicitar en qué sentido los hallazgos científicos resultan relevantes para un enfoque filosófico de la mente y el cuerpo, y a ubicar la realidad de los fenómenos mentales dentro de ese marco. (Rabossi 1995: 21-22)

No es de extrañar, entonces, que cuando la psicología cognitiva comenzara a abundar en la postulación de representaciones internas y reglas a las que obedecía la manipulación de dichas representaciones, la legitimidad de tal postulación –sobre todo, la posibilidad de anclarla en una concepción naturalista de la realidad– ocupara pronto las primeras líneas de la refriega filosófica, concediendo así a las preocupaciones de Brentano (1874, *supra*) una primacía que habría repugnado a cualquier conductista de bien. Tampoco resulta raro que el apogeo del cognitivismo trajera consigo una vivísima fascinación por el problema de la consciencia –por sus tentativas de naturalización y su eficacia causal, fundamentalmente–, que ya Watson, en su campaña contra el introspeccionismo y el estructuralismo, había repudiado sonoramente al declarar que "[…t]he time seems to have come when psychology must discard all reference to consciousness" (Watson 1913: 163).

También en el estudio de García-Carpintero (1995) acerca del funcionalismo éste queda perfilado "[…] como un desarrollo de temas enfatizados por el conductismo [lógico], resultado de considerar objeciones a esta doctrina" (García-Carpintero 1995: 54). Las dificultades a las que Ryle y los suyos no habrían logrado sobreponerse, y que habrían impulsado el ascenso del funcionalismo, serían en esencia dos: que los estados mentales no parecen venir constituidos por las disposiciones conductuales en que se manifiestan, sino más bien constituir la causa, o acaso el fundamento categórico, de dichas disposiciones (*cf*. Putnam 1967a: 230, *infra*) y que "[…] el concepto de un estado mental parece estar inextricablemente unido al de otros estados mentales, de modo que los estados mentales forman un todo interrelacionado" (García-Carpintero 1995: 53).

La relación entre ciertos estados psicológicos –las emociones– y las alteraciones fisiológicas o conductuales que característicamente los acompañan es un asunto en tornoo al cual ya William James había sembrado la sospecha en 1884: tal vez esta afectación fisiológica o conductual no fuera, como solemos pensar, la expresión de los cambios afectivos, sino más bien su antecedente; tal vez lo cierto –insistía James, como es bien sabido– resulte ser "[…] que nos sentimos tristes porque lloramos, enfadados porque golpeamos, asustados porque temblamos […]" (James 1884: 59), y no al contrario. Con el propósito de resguardar su hipótesis de la "[…] inmediata incredulidad […]" que anticipa que despertará así esbozada, James nos invita poco después a un breve ejercicio de imaginación:

> Si nos imaginamos una emoción intensa y tratamos entonces de abstraer de nuestra consciencia de ella todas las sensaciones de sus síntomas corporales característicos, nos encontramos con que no nos queda nada, no hay un "ingrediente mental" a partir del cual pueda construirse la emoción, de modo que todo lo que queda es un estado de percepción intelectual, frío y neutro. (James 1884: 62)

No es difícil ver, por una parte, que una concepción netamente conductista de la emoción resultaría de la sencilla maniobra de prescindir del "estado de percepción intelectual" descrito en la última cláusula de la oración de James, algo a lo que el propio James parece invitarnos al adelantar justo antes de mencionarlo la

inexistencia de un "ingrediente mental a partir del cual pueda construirse la emoción". Que los vaivenes retóricos de James resulten confusos –de *no nos queda nada* a *no nos queda nada de naturaleza mental*, y de ahí a *lo que nos queda es un estado perceptivo de naturaleza intelectual*– seguramente no sea más que un pecado venial que el ardor de la controversia haría excusable; que en el pensamiento de James arraiga en buena medida la tradición conductista norteamericana es poco más que un lugar común. Pero vale la pena señalar que el ejercicio de imaginación que James despacha sin muchos más escrúpulos regresaría tanto tiempo después, casi inalterado, precisamente para minar la credibilidad de un ya debilitado programa conductista. Bien es cierto que ya James dejó constatado que aunque "[…] la mayoría de la gente […] afirma que su introspección verifica esta afirmación" –interpretándola en el sentido más moderado: la de que si imaginamos una emoción desprovista de sus síntomas corporales lo que imaginamos es sólo un estado intelectual–, "algunos se obstinan en afirmar que la suya no" (James 1884: 62). Pero cuando a renglón seguido James culpa de esa obstinación a una deficiente comprensión del problema, parece que su expeditivo dictamen pudo ser cuando menos algo precipitado[61].

En efecto, si Putnam (1963a) pudo orquestar los reinos imaginarios de los superespartanos –que ocultan su dolor a la perfección sin dejar de sentirlo[62]– y los superactores –que fingen a la perfección un dolor que no llegan a sentir–, convencernos con cierta facilidad de que las disposiciones conductuales características de un determinado estado mental pueden darse en ausencia de éste tanto como éste en ausencia de ellas, y, de esa manera, "enterrar el conductismo lógico" (Putnam 1963a: 327)[63], es sin duda porque la obstinación de quienes insistían en poder imaginar una emoción desprovista de sus expresiones corporales habituales era más pertinaz, o por lo menos más refractaria a las buenas razones de James, de lo que nos habría sido dado esperar si su fundamento hubiera radicado sólo en un pobre discernimiento del asunto. La convicción que James (1884: 62) hizo cristalizar en la tesis de que "[…u]na emoción humana puramente incorpórea es un ente vacío" bien pudo leerse como una velada exhortación al conductismo, o al menos un paso

[61] En unas conocidas líneas de las *Investigaciones* habría de preguntarse Wittgenstein (1953: §281): "¿Pero no se reduce lo que tú dices a esto: que no hay dolor, por ejemplo, sin *conducta expresiva de dolor*?" Su respuesta es negativa: "Se reduce a esto: sólo de un ser humano vivo y de lo que es semejante (se comporta de un modo semejante) puede decirse que tenga sensaciones". El giro "puede decirse" apunta a lo que acaso podría llamarse un *conductismo gramatical*, evocando lo que poco después contestaría, esquivo, el propio Wittgenstein (1953: §307) tras preguntarse "¿No eres tú entonces un conductista disfrazado? ¿No dices tú, en el fondo, que todo es ficción excepto la conducta humana?" –a saber: "Si hablo de una ficción, entonces es de una ficción *gramatical*". Idéntica lectura del problema del conductismo en Wittgenstein puede hallarse en Hierro-Pescador (2002: 57-59).

[62] En realidad, quienes así se comportan serían los super-superespartanos, pues los meros superespartanos, en el experimento imaginario de Putnam (1963a), conservan al menos formas de expresión verbal del dolor asimilables a las nuestras.

[63] No deja de resultar curioso que, años después, Searle (1992: *xi*) recurriera a casi la misma metáfora para referirse a su propio e influyente intento de refutar irrevocablemente el funcionalismo: "I want to put the final nail in the coffin of the theory that the mind is a computer program".

en la dirección en la que Watson caminaría decididamente, aun cuando lo que hubiera querido decir James no fuese que una emoción humana puramente incorpórea no es nada, sino que es sólo un estado intelectual y no la pretendida emoción. Sin embargo, esa misma convicción quedaría inerme cuando, desaborlado ya el conductismo, se enfrentara a la intuición opuesta de que nuestra vida emocional puede pervivir en ausencia de sus signos externos. La razón, si la crítica cognitivista del conductismo no ha sido errada, sería justamente que las disposiciones conductuales –o los signos externos– no *conforman* el estado mental, sino que *son su efecto* en un rango significativo de circunstancias[64].

Dicho de otro modo: no puede trasladarse al conductismo lógico la réplica de Hempel (1935) al argumento del actor perfecto, que ya se anticipaba en aquel trabajo. Según Hempel (1935: 19), la hipótesis de que un estado mental pudiera darse en ausencia de sus "síntomas físicos" es una contradicción lógica equivalente a la de asegurar que una afirmación puede ser falsa al tiempo que se verifican las condiciones necesarias y suficientes de su verdad. Pero si por "síntomas físicos" se entiende en el argumento de Hempel lo que necesitaría entender un conductista lógico –es decir, conductas o disposiciones a la conducta–, entonces la réplica no es más que una abierta petición de principio, puesto que la tesis que Hempel trata de establecer es precisamente la de que las condiciones de verdad de los enunciados mediante los cuales atribuimos estados mentales aluden a sus manifestaciones comúnmente observables, es decir, conductuales. En cambio, si, quizá con mayor fidelidad hacia la posición del propio Hempel, se entiende por "síntomas físicos" cualquier alteración física del organismo, ya sea de naturaleza conductual o fisiológica, entonces, aparte de las dificultades que entraña remitir las condiciones de

[64] Dicho sea de paso: la hipótesis de que las alteraciones afectivas sea *en condiciones normales* posterior a los cambios fisiológicos manifiestos que las acompañan es una interpretación mínima de la posición de James que es fácil hacer compatible con los argumentos de Putnam. En efecto, nada en la concepción funcionalista de los estados mentales parece impedir que algunos de ellos vengan causados por determinados estímulos o estados internos, causen determinadas respuestas o estados internos –siendo así que ese patrón de relaciones funcionales delimita la clase de estados mentales en cuestión–, y además resulten ir precedidos, al menos en sujetos humanos y *en condiciones normales*, de determinadas alteraciones fisiológicas manifiestas –que, sin embargo, no forman parte del patrón de relaciones funcionales que deslinda su naturaleza como clase. Incluso más: la tesis funcionalista de Putnam podría asimilar sin resquebrajarse que esas alteraciones fisiológicas manifiestas posean, como intuye James, un papel causal en la aparición de las emociones en humanos, siempre y cuando (*i*) quepa caracterizarlas como instancias de una clase más amplia de estímulos que provocan emoción –en humanos o en otros sistemas dotados de vida emocional y, por tanto, probablemente en términos que no sean propiamente fisiológicos–, y (*ii*) no se trate del único tipo de estímulos o estados internos que poseen tal papel causal –o dicho de otro modo, no constituyan condición necesaria para la aparición de la emoción–, dejando así margen para la intuición de que los super-super-espartanos no son una ficción inconcebible. Huelga decir, en todo caso, que ya la mera descripción de las comunidades imaginarias bosquejadas por Putnam parte de una perspectiva opuesta a la de James, en la medida en que asume la noción de *expresión* de las emociones –lo que los super-super-espartanos retienen y los super-actores remedan–, que es precisamente la noción que James provocativamente cuestiona.

verdad de enunciados cotidianos –y con ello, según Hempel, su significado– a hechos que no son siquiera observables en el contexto cotidiano, nos encontraremos con que interpretar ese argumento como una defensa del conductismo exige forzar la noción de conducta –"conducta corporal hasta el mínimo detalle", dice Hempel (1935: 19) tratando de abarcar conducta y actividad fisiológica bajo un mismo concepto– en un sentido tal que la absorción del conductismo por alguna variedad de reduccionismo psicofisiológico se torna inevitable. Bajo este prisma, además, el argumento simplemente ignora las intuiciones medulares del funcionalismo: que la actividad fisiológica en la que consista un determinado estado mental bien puede ser distinta en distintos casos de ese mismo tipo de estado mental, que los estados mentales son estados funcionales cuya encarnación fisiológica puede ser sumamente variable, que esos estados funcionales son causas de la conducta observable[65].

La confusión entre efectos y constituyentes de lo mental sería pues, irónicamente, el error categorial sobre el que se fundaría el conductismo lógico –tal como el dualismo cartesiano, según el diagnóstico del propio Ryle (1949), se habría erigido sobre la confusión entre cosas y disposiciones. Así, resulta que los efectos de un determinado estado mental pueden serlo de otras causas si las circunstancias son otras –las conductas típicamente asociadas al dolor, por ejemplo, serlo de las portentosas dotes interpretativas de los superactores–, al igual que puede la causa darse sin el efecto –la sensación de dolor desprovista de todo residuo de sus signos habituales por obra de la férrea voluntad de los super-super-espartanos, o bien darse con efectos distintos, en circunstancias distintas. Todo ello resultaría impensable si los efectos del dolor, sus signos, fuesen como quisiera Ryle constituyentes del dolor mismo.

En último término –como apuntaría el propio Putnam (1967a)– pesa contra el conductismo lógico un mera cuestión de credibilidad intuitiva. Incluso si el conductista llegara a proporcionar un análisis del estado de sentir dolor que no mencionara otros estados mentales –o que, piensa Putnam (1967a: 229, *infra*), no terminara a fin de cuentas mencionando el propio estado de dolor sometido al análisis–, e incluso si ese análisis pudiera sobreponerse de algún modo a las diferencias entre especies, o entre organismos y máquinas, que condensadas en la noción de realizabilidad múltiple vertebrarán el ataque de Putnam (1967a) a la tesis de identidad psicofísica y su defensa del funcionalismo, incluso en tan improbables circunstancias, los parapetos del conductista lógico ante el empuje de una concepción funcional de la mente (o, a estos efectos, de una concepción fisicalista) seguirían siendo precarios. Así,

[65] Lo mismo ocurre, *mutatis mutandis*, si intentamos forzar la lectura del argumento de James (1884) para que las referencias a las alteraciones fisiológicas manifiestas que acompañan a la emoción se tomen como referencias a alteraciones fisiológicas *tout court*, incluidas las no observables. Si bien algunos pasajes del influyente artículo publicado en *Mind* pueden instigar esa interpretación (*cf.* por ejemplo, las observaciones de James (1884:60) sobre el pletismógrafo), es patente que ello, si pretendiese convertirse en una defensa del conductismo lógico, sólo conduciría al mismo atolladero al que nos aboca la misma táctica en el caso de Hempel (1935).

> […] it would still be more plausible to identify being in pain with some state whose presence *explains* this behavior disposition –the brain state or functional state– than with the behavior disposition itself. (Putnam 1967a: 230)

Naturalmente, a esas consideraciones de plausibilidad –que Putnam (1967a: 230) mismo no duda en tildar de "subjetivas", y que parecen extraer parte de su pujanza de lo que Searle (1992: 35, *infra*) da en llamar la "objeción de sentido común" contra el conductismo– bien podría el conductista oponer otras de parsimonia: ¿qué motivo podría haber, en las circunstancias descritas, para diferenciar, además de la conducta, la disposición conductual y el estado físico o funcional que explica dicha disposición, en lugar de conformarnos con la conducta y la disposición a que ésta se dé? Dirimir si es la parsimonia o la credibilidad intuitiva lo que debe primar en este asunto requiere, sin duda, un cuidadoso análisis de la noción de disposición sobre la que orbita la controversia.

Antes de abordar dicho análisis, merece la pena apuntar que la idea de que la relación entre estados mentales, de un lado, y estímulos y conductas –o disposiciones–, de otro, pertenece al ámbito de la lógica ha sido esgrimida por el conductismo filosófico para tratar de decretar una refutación *a priori* de las tesis funcionalistas –así lo ha hecho, por ejemplo, Malcolm (1984). Es innegable que *si* las relaciones entre, por ejemplo, la sed y la ingesta de líquido son relaciones lógicas, *entonces* es de rigor acometer una profunda revisión del planteamiento fundamental del funcionalismo según el cual un estado mental, como la sed, viene identificado por las relaciones causales que establece con diversas conductas, como la ingesta de líquidos, estímulos, y otros estados internos. Al alcance del funcionalista sólo queda entonces, como bien supo ver Bechtel (1988: 165), cuestionar que las relaciones entre estímulos y respuestas, por una parte, y estados mentales, por otra, sean efectivamente de orden lógico. La tarea, en cualquier caso, parece asequible, dado que es un lugar común del propio discurso cotidiano sobre lo mental la poca mansedumbre de los conceptos que lo conforman –la tozuda propensión de creencias, deseos o sentimientos a desmanarse, a irrumpir en circunstancias que no les corresponden o provocar inesperadas consecuencias–, y dado, sobre todo, que los análisis conductistas poco han podido hacer por domeñar estas tercas nociones, como hace ver precisamente la ubicuidad de las cláusulas *caeteris paribus* que los adornan. Renunciar a la idea de unas límpidas relaciones lógicas entre estímulos, conductas y estados mentales y reemplazarla por la caracterización de estos últimos en virtud de su lugar en una maraña de causas y efectos que los liga a estímulos y conductas parece, al menos a primera vista, una buena manera de dar cuenta del indómito carácter de nuestros conceptos coloquiales referidos a estados mentales. Como anota Bechtel:

> Un estado mental estará conectado con una gran variedad de estados mentales diferentes en esta red causal. Esto abre la posibilidad de que seamos capaces de identificar el estado mental en una gran variedad de maneras diferentes. Cualquiera de estas maneras puede

> considerarse como falible, revisable si otras maneras de identificar estados mentales nos llevan a una conclusión diferente. Cuando diversos indicadores diferentes señalan en la misma dirección, entonces nuestra evidencia a favor del estado mental es más fuerte y más fiable que cuando debemos fiarnos de un solo indicador (Bechtel 1988: 165).

De todos modos, la debilidad de la noción de disposición que operaba en los análisis de Ryle (1949) había sido ya duramente reprobada por Peter T. Geach (1957). Irritaba sobremanera la sensibilidad ontológica de Geach que el devenir de nuestra vida mental pudiera consistir en una sucesión de fenómenos de índole meramente disposicional, carentes de todo fundamento categórico: esa era la posición que defendía Henry H. Price (1953: 322, *apud* Armstrong 1968: 86), compañero de Ryle en Oxford, así como la que parecían implicar los argumentos del propio Ryle, según él mismo reconocería informalmente ante el profesor Charles B. Martin, de quien Place (1999: 385) lo habría escuchado en persona en 1995[66]. Con una ironía igualmente fina, Geach repudiaba expeditivamente una perspectiva que consideraba insólita en el contexto del pensamiento científico[67]:

> Of course there may be people prepared to say that, although men of science regularly look for differences already existing between the agents in order to explain differences of behavior, there is no reason to expect that such differences always do exist; the principle on which men of science proceed might be as unsound as a gambling system, and their success up to now mere luck. I shall not argue the point. (Geach 1957: 6)

Resulta a duras penas discutible que entender –pongamos por caso– la fragilidad de un cristal como su tendencia a quebrarse en determinadas circunstancias no es óbice para atribuir esa disposición a ciertas propiedades del cristal –de su estructura molecular– que constituyen la base categórica de la disposición[68]. Como hiciera notar Armstrong (1968: 86),

> […] in asserting that a certain piece of glass is brittle […] we are ipso facto asserting that it is in a certain non-dispositional state which disposes it to shatter and fly apart in a variety of circumstances[69].

Ésa era, de hecho, la concepción de las disposiciones de Carnap (1932, 1938), de la que Ryle parece partir pero se aparta en este punto[70]. Que un movimiento parejo al

---

[66] Tanto Place (1999) como Heil (1989a) y, mucho antes, Presley (1967) han tratado de subrayar la importancia de las aportaciones de Martin en la génesis de la tesis de identidad psicofísica, que quedaría ligada sin embargo a los nombres de Smart, Lewis y el propio Place.

[67] No es difícil, desde luego, adivinar entre líneas la reprobación que probablemente despertaran en Geach, un tomista, los bien conocidos intereses de Price en el ámbito de la parapsicología, que lo habían llevado a presidir la *Society for Psychical Research* de Londres entre 1939 y 1940.

[68] Ni Quine (1975: 87) vacilaría en asegurar que "[…a] disposition is in my view simply a physical trait, a configuration or mechanism […]. Dispositions to behavior, then, are physiological states or traits or mechanisms".

[69] Pero *cf*. Squires (1968) para una ingeniosa objeción de circularidad, que repasaremos someramente *infra*, contra el análisis de Armstrong.

ensayado por Carnap estuviera vedado a la teorización psicológica, so pena de incurrir en las explicaciones "paramecánicas" denostadas por Ryle, es contra lo que los funcionalistas se rebelarían –un detenido examen de los argumentos contra Ryle en Fodor (1968) es iluminador a este respecto.

En defensa del planteamiento de Ryle, sin embargo, aduce Place (1999) que, aun cuando Ryle coqueteara con la posición disposicionalista radical de Price (1953), y aun cuando su insensibilidad a las preocupaciones y los patrones explicativos propios de la investigación científica fuese patente, es injusto acusarle de pretender que las disposiciones sean realidades últimas e inexplicables, o de intentar ocultar su mera ignorancia sobre las causas eficientes de la rotura de un cristal. El empeño de Ryle –dice Place (1999: 384)– es el de dar cuenta del significado de los enunciados disposicionales, pero lo que quiera que explique los hechos que vienen descritos por un enunciado disposicional "[…] is a scienfic matter and, as such, none of Ryle's business".

Ahora bien: ni Ryle hace honor en sus juicios sobre la explicación psicológica a esa respetuosa indiferencia respecto a los quehaceres científicos –de ahí que despertara la minuciosa respuesta de Fodor (1968)–, ni el propio Place es ajeno en su defensa de Ryle a cierto puritanismo conceptual sobre lo que constituye una explicación adecuada y lo que no. En efecto, Place (1999: 383-384) intenta desbaratar la grave acusación de Geach (1957) según la cual Ryle habría incurrido en una apología de la pseudoexplicación tautológica, al estilo de la pretendida explicación de los efectos sedantes del opio por una *virtus dormitiva* que éste albergara, o de la *vis prolifica* que a juicio de los ignorantes médicos parodiados por Molière ocasionaría la mayor o menor facilidad para engendrar descendencia, o –lo que es peor– de algunas de las explicaciones psicológicas de las que el mismo Ryle, como Skinner, se burlaría (*cf. infra*).

Explicar –dice Place– la capacidad del opio de sedar a los pacientes como efecto de su *virtus dormitiva* es tautológico porque el *explanandum* es tan general como el *explanans*. Ciertamente, es a la pregunta del doctor por las "causam et rationem quare opium / Facit dormire" que el bachiller replica en su latín macarrónico: "A quoi respondeo, / Quia est in eo / Virtus dormitiva / Cuius est natura / Sensus assoupire" (Molière 1673: III). Cuando Ryle explica la rotura de un cristal como efecto de su fragilidad, en cambio, el *explanandum* es un evento particular y el *explanans* un enunciado disposicional legaliforme, capaz de sustentar contrafácticos (Place 1999: 383-384): la analogía con la *virtus dormitiva* se quiebra y la apariencia de circularidad se diluye. De modo que, según argumenta Place:

> If what was to be explained *had been* a particular case of someone going to sleep after taking opium, an explanation that attributed that concatenation of events to opium's

[70] Aunque Ryle no lo cita, su elección de la fragilidad y la solubilidad a modo de ejemplos de disposición evoca nítidamente a Carnap (1932: 186; 1938: 53). Cómo Ryle se aleja de Carnap en la cuestión crucial del fundamento categórico de tales disposiciones ha sido destacado entre otros por Medlin (1967: 115); de todo ello toma nota también Pujadas (2002: 21).

> *virtus dormitiva* […] would not only have been a perfectly acceptable explanation; it would be the only kind of explanation required in such a case, an explanation that subsumes it under the relevant law. (Place 1999: 384)

La observación de que mencionar la estructura molecular del cristal o del opio no procede, *sensu stricto*, en la explicación de por qué se rompe un cristal determinado o por qué un paciente se adormece bajo el efecto del narcótico, sino en la de la fragilidad del cristal o el efecto sedante del opio en general, es tan conceptualmente sólida –bajo los presupuestos sobre la idea de explicación que operan en Place– como pragmáticamente endeble. Porque cuando inquirimos por la explicación de un evento particular, la mención de la ley o del enunciado legaliforme bajo el cual el evento resulta explicable no calma la inquietud que originó la pregunta. Lo que querríamos saber es, por lo común, cuáles son las causas del evento que nos interesa, y *bajo la idea de causalidad que alberga el sentido común* que un suceso cualquiera quede subsumido en una generalización no es la causa de que suceda. La explicación no nos resulta convincente hasta que averiguamos qué mantiene en pie la generalización, por qué se verifica: es decir, hasta que nos vemos en disposición –aunque no siempre precisemos hacerlo– de detallar la aplicación al caso particular que nos interesara de las secuencias causales que con carácter general respaldan la ley que lo cubre. En realidad, si pregunto en la vida cotidiana por qué se ha roto un cristal, lo más seguro es que la respuesta que espere sea simplemente qué o quién lo ha golpeado: "porque es frágil" no responde a mi pregunta. Pero si pregunto con algún ánimo de indagación científica por qué se ha roto un cristal, lo más seguro es que la respuesta que espere esté modelada bajo la misma noción de causalidad eficiente que trasluce en la cuestión de qué o quién lo ha golpeado, y que no ataña a la generalización según la cual el cristal, en condiciones como las que han afectado al cristal particular que ha motivado mi curiosidad, se rompe, sino a la causa singular que lo ha roto en el caso que me interesa: sea como sea, "porque es frágil", de nuevo, no responde a mi pregunta. En suma: dado que la noción de causalidad resulta después de todo inseparable de nuestra idea de explicación de un evento –o al menos alguna noción de causalidad lo es bajo al menos algunas de las acepciones de "explicar"–, pretender contener la descripción de las estructuras o mecanismos que den cuenta de una generalización hasta que haya una pregunta explícita por los fundamentos de la generalización es, bajo las antedichas acotaciones, ignorar flagrantemente las expectativas de quien pregunta por qué tuvo lugar el evento en cuestión. Así pues, al dictaminar que enunciar la ley general que lo acoge es *la única explicación relevante* de un evento particular, Place parece, a fin de cuentas, embarcado de pronto en el mismo empeño de poner coto a la explicación científica de las disposiciones que él rechaza abiertamente, y que no sin cierta condescendencia atribuye (Place 1999: 384) a la formación en letras clásicas de un Ryle cuya entera carrera académica estuvo vinculada a Oxford. Que ese empeño fuera combatido con determinación en el seno

de un incipiente cognitivismo, dada la notoriedad que había alcanzado el trabajo de Ryle, era tan inevitable como pueda serlo cualquier acontecimiento histórico[71].

No menos delicadamente hiriente para con Ryle parece la posición de Geach respecto de la lógica del análisis disposicional –un flagrante ejemplo, a su juicio, de explicación *obscurum per obscurius* (*cf.* Arnauld y Nicole 1662/1759: 297, *supra*):

> It ought to be, but plainly is not, generally known to philosophers that the logic of counterfactual conditionals is a very ill-explored territory; no adequate formal logic for them has yet been devised, and there is an extensive literature on the thorny problems that crop up. It is really a scandal that people should count it as a philosophical advance to adopt a programme of analysing ostensible categoricals into unfulfilled conditionals, like the programmes of […] neobehaviourists with regard to psychological statements. (Geach 1957: 6-7)

También el criterio de Putnam (1975b) se hizo oír en la controversia sobre el papel explicativo de las disposiciones y de sus fundamentos categóricos, aunque de forma que acaso pueda verse como irónica. El ejemplo esgrimido por Putnam no podía ser más sencillo: si tenemos un cubo metálico de $^{15}/_{16}$ pulgadas de lado, una abertura redonda de una pulgada de diámetro, y una abertura cuadrada de una pulgada de lado, practicadas ambas en una tabla de madera, podemos explicar fácilmente el hecho de que el cubo pase por la abertura cuadrada pero no por la redonda en términos de la rigidez de ambos sólidos y de sus propiedades geométricas. También podría, claro, ensayarse una explicación en términos de las estructuras atómicas del cubo y la tabla, de tal modo que fueran las leyes de la mecánica de partículas las que, junto con la información detallada sobre dichas estructuras y sus trayectorias posibles, condujera a la conclusión de que no es posible que el cubo atraviese la abertura redonda, y sí la cuadrada. Pero, a juicio de Putnam, esa complejísima deducción de la imposibilidad de una trayectoria del cubo y la posibilidad de otra no cuenta como una explicación –o por lo menos no como una buena explicación– porque adolece de la necesaria generalidad: no sólo no trae a colación los hechos relevantes para comprender el fenómeno, sino que los oculta en un maremágnum de

---

[71] Cuestión bien distinta es –cabe anticipar– que en *toda* relación causal haya de existir de hecho un mecanismo eficiente que, presumiblemente en el nivel de descripción de los constituyentes de la materia, pueda dar cuenta de cada instancia particular de concatenación de causa y efecto al margen de las regularidades que al respecto puedan quedar apresadas en distintas leyes o generalizaciones legaliformes, o que, al contrario, exista un sentido legítimo de la noción de causalidad –un sentido humeano– en el que ésta se reduzca al entramado nomológico en el que una determinada propiedad se encardine, o, desde luego, que *toda* relación causal resulte no ser sino una suerte de imbricación nomológica para la que no quepa identificar en último término un sustrato mecánico –o, si se quiere, paramecánico.

Las conclusiones a las que nos ha de abocar este estudio incumben a la imposibilidad de afianzar definitivamente la autonomía de la explicación psicológica, más allá de una cierta idea de relevancia explicativa fundada en consideraciones pragmáticas, sin tomar posición en el debate sobre la metafísica de la causalidad., pues tal afianzamiento exige articular el espacio ontológico en que lo mental pueda atesorar su propio vigor causal.

datos y cómputos sobre estructuras atómicas. Trataba Putnam, de ese modo, de aliviar el temor a que la admisión del materialismo –cuanto existe es materia– nos arrastre a admitir también la verdad del fisicalismo –cuanto existe tiene una explicación física–, temor que habría atenazado tanto al lego como –dice Putnam (1975b: 295)– a Descartes o Diderot, pero que carece de fundamento: de igual modo que la explicación geométrica es autónoma respecto de la deducción microfísica, también lo es la explicación psicológica. De qué estemos hechos no es tan importante al fin y al cabo: en la que sería una de las formulaciones más osadas y jocosas del desdén hacia la explicación neurofisiológica en que a veces tornó la afirmación de la autonomía de la psicología, Putnam (1975b: 291) abría su trabajo asegurando que "[..w]e could be made of Swiss cheese and it wouldn't matter"[72].

Años después, Block (1997) recordaría la aportación de Putnam con matices muy retocados. Lo que Putnam nos habría enseñado –parece pensar Block– es que la explicación disposicional es más general, pues puede aplicarse a cuerpos sólidos con estructuras atómicas diferentes, como cristal o madera, mientras que la explicación categórica tiene el valor crucial de que permite dar cuenta de las propiedades que aparecen en la explicación disposicional, como la solidez o la rigidez, y entender los detalles de su naturaleza. Así, Putnam habría acertado a delimitar lo provechoso del ensayo conductista de caracterización relacional de lo mental –queda reivindicada la importancia de un nivel de explicación funcional, al que apuntaba parcialmente el análisis disposicional de Ryle–, a la vez que lo encajaría con la insistencia del fisicalismo en la eficacia causal de los estados mentales –idénticos, según la doctrina fisicalista, a estados cerebrales–, desautorizando la insistencia de Ryle en ceñirse a lo disposicional. Con ello –cabe añadir– su juicio se anticipaba al que habría de formarse Fodor (1981a: 9, *infra*) sobre la capacidad del funcionalismo de conjugar ambas intuiciones. Pero –ya lo hemos visto– la posición de Putnam (1975b) no era exactamente ésa, y seguramente contribuyera a dar pábulo a la preocupación por el talante excesivamente antifisiológico de la entonces incipiente ortodoxia funcionalista, que saltaría a la vista también en Marr (1982: 36, *infra*).

No en vano, todavía Elliott Sober (1999), un discípulo del propio Putnam, intentaría contribuir al desmantelamiento de la interpretación antirreducionista del funcionalismo inaugurada por Putnam (1967a) poniendo en tela de juicio la asunción, que achaca a su maestro, de que la mayor generalidad de una explicación es siempre un criterio decisivo para preferirla. A juicio de Sober (1999: 550), en cambio, es importante tener en cuenta que la ciencia está tan comprometida con la amplitud como con la profundidad de sus explicaciones. Así, en el terreno de la

[72] Así pues, si lograba proveerse de una reconstrucción disposicional de las relaciones geométricas mencionadas por Putnam, el conductista lógico tendría al alcance de la mano un robusto parapeto tras el que resguardar su desinterés por el fundamento categórico de las disposiciones en las que consisten a su entender los estados mentales, remedando el argumento de Putnam. No era la primera vez que los decididos argumentos de Putnam contra el fisicalismo mutaban a su pesar en una posible defensa del conductismo: ya en un trabajo seminal como Putnam (1967a) se había dado, como se analizará detenidamente *infra*, un requiebro parecido.

disputada reducción psicofísica, la explicación categórica de las causas de una conducta concreta, o un estado psicológico concreto, mediante la mención de los procesos fisiológicos o físicos que constituyen dichas causas –y, se entiende, también de los que constituyen el efecto: la conducta o el estado mental en sí– es un objetivo tan loable en términos epistemológicos como el de subsumir ese par causa-efecto en generalizaciones de la mayor envergadura posible, sean éstas o no de índole funcional o disposicional. Esto –anota Sober (1999: 562)– es al fin y al cabo una cierta variedad de reduccionismo, y no sólo no colisiona con el hecho de que estados psicológicos del mismo tipo puedan encarnarse en estados física o fisiológicamente muy dispares, sino que se sigue precisamente de ese hecho.

Con todo, conviene constatar que incluso Place, acaso la voz que con más autoridad persevera en la restitución de los logros teóricos de Ryle, acepta que la estrategia de tratar tanto los enunciados nomológicos como los que atribuyen disposiciones como permisos de inferencia –"inference licences" o "inference tickets" dirá, *passim*, Ryle– es fruto de una deficiente concepción de la relación entre la forma condicional de un enunciado y la atribución de causalidad que tal enunciado puede o no comportar. La confusión que Place denuncia en el planteamiento de Ryle consiste –si se quiere– en amalgamar una modalidad *de re* y una modalidad *de dicto* en la atribución de relaciones condicionales, cuando sólo la modalidad *de re* conllevaría atribución de causalidad. Se trata, en palabras de Place, de:

> […] a failure to distinguish between, on the one hand, causal conditionals that describe conditional relations between the existence or occurrence of affairs and events and, on the other, sentences of the form 'If *p*, then *q*" that describe conditional relations between the truth of propositions. (Place 1999: 379)

Que esa no es una distinción sin diferencia queda claro, a juicio de Place, si consideramos el argumento desplegado al respecto por Martin (1994), cuyo efecto sería devastador para el análisis disposicional de Ryle si por enunciado disposicional se entiende una descripción de la relación entre los valores de verdad de varias proposiciones –nada más que un permiso de inferencia–, pero del todo inocuo si Ryle hubiera sabido ver los enunciados disposicionales como generalizaciones legaliformes sobre las relaciones entre hechos mundanos. El intento emprendido por Martin de refutar el análisis ryleano parte de un ingenioso artefacto imaginario –el "electro-fink": algo así como "electro-chivato"– capaz de detectar la conexión entre un cable de fase y un borne de tierra y, a nuestro antojo, eliminar la carga del cable de fase tan pronto como se dé la conexión, impidiendo así que se produzca corriente, o, si hubiéramos eliminado la carga del cable previamente, cargarlo a la par que se produce la conexión, posibilitando esta vez la corriente. Una breve consideración del funcionamiento del artilugio hace patente que la equivalencia entre "Este cable es de fase" y "Si este cable hace contacto con un borne de tierra se produce una corriente eléctrica de aquél a éste" es insostenible: la corriente puede darse sin que el cable fuera de fase antes de producirse el contacto, si el invento lo detecta y carga el cable,

o puede no producirse aunque el cable fuera de fase, si el invento detecta el contacto y neutraliza el cable. Por supuesto, la equivalencia desarticulada por Martin es un ejemplo relativamente sencillo del tipo de análisis disposicional favorito de Ryle. Pero aquí es donde interviene el *distingo* que Place pronuncia. Si los enunciados disposicionales de Ryle se reconstruyen –como él mismo insistía en que se hiciera– como permisos inferenciales del mismo orden que $p \rightarrow q$, entonces Martin, en efecto, ha logrado desmantelarlos. Pero si, en cambio, damos en entenderlos como generalizaciones sobre relaciones causales entre eventos, con soporte contrafáctico pero, igual que cualquier generalización de esa índole, sometidas a cláusulas *caeteris paribus*, el problema se diluye. Sencillamente –dice Place– entre las cosas que las cláusulas *caeteris paribus* del análisis disposicional de "Este cable es de fase" tendría que descartar se cuenta la posibilidad de que el cable esté conectado a un dispositivo como el imaginado por Martin, o cualquier otro que interfiera "[…] between a disposition and its manifestations in such a way as to create the disposition whenever it would otherwise *not* exist or remove it whenever it *would* otherwise exist" (Place 1999: 395).

No obstante, incluso si aducimos en defensa de Ryle un argumento al que cabe sospechar que hubiera mostrado reticencias, como sin duda es legítimo hacer, lo que Place logra no es más que esquivar el escollo detectado por Martin a costa de adentrarse en los arrecifes de la imprevisibilidad de las cláusulas *caeteris paribus*, cuyo farallón más amenazante para los intereses de Ryle no es otro que el problema del círculo de lo mental. Es esa imprevisibilidad, en efecto, lo que que nos obliga a dar a las cláusulas *caeteris paribus* de los análisis ryleanos una formulación inaceptablemente vaga o, peor aun, circular, que salta a la vista en el ejemplo del cable de fase con tanta crudeza como, según veremos, en el de la creencia de que va a llover. La peculiaridad del caso del análisis disposicional de enunciados referidos a estados mentales reside sólo, así visto, en que se había estipulado de entrada, precisamente, que los enunciados referidos a estados mentales que recurren en las cláusulas *caeteris paribus* iban a ser traducidos sin residuo a términos estrictamente disposicionales. De cara al análisis disposicional de "Este cable es de fase", por el contrario, no está en vilo la promesa de que el análisis quede depurado de toda referencia a fenómenos eléctricos. En definitiva, la defensa de Ryle ante Martin ensayada por Place es firme, pero hace descollar otros de los problemas de la concepción ryleana de lo mental, que seguramente sean infranqueables.

Sea como sea, parece claro que la táctica de Place evita cuidadosamente, como Ryle, dotar a las disposiciones (*ergo*, si admitimos sus tesis, a las actitudes proposicionales) de toda eficacia causal. Los enunciados disposicionales, tal como los interpreta Place, expresan relaciones causales entre eventos, pero la relación de una disposición con un evento concreto no es la de causa con efecto. Antes bien, la disposición es a las estructuras, mecanismos o procesos que sostienen la relación causal entre eventos particulares como el efecto a la causa: la disposición es el efecto de esas estructuras, si bien bajo el prisma de una generalización legaliforme (Place 1999: 393, *infra*). Pues bien, cabe ver el funcionalismo, en buena medida, como un

intento de enrumbar nuestro concepto de lo mental hacia la eficacia causal que el conductismo lógico le había denegado al desproveerlo de fundamento categorial, y hacerlo sin volver a zozobrar en los bajíos en que habría encallado el dualismo cartesiano, y que Ryle había cartografiado con mano firme. El norte de esa carta de navegación es lo que habría proporcionado la extensión de la teoría de la identidad psicofísica operada por Armstrong (1968), al igualar estados mentales y estados cerebrales y dotar así a los primeros de la intachable naturaleza causal de los segundos (*cf.* Fodor 1981a: 9, *infra*). El terreno intermedio, en el que cabe conceder tanto existencia *hic et nunc* como eficacia causal a sensaciones o imágenes mentales pero no a creencias o deseos, sería el habitado por Place (1956).

El (retorno del) problema del retorno de lo mental

El análisis de la respuesta de Place (1999) a Martin (1994) ha revelado que la relectura del significado de los enunciados disposicionales que Place propone no logra, aunque acaso tampoco lo pretenda, que la concepción de las actitudes proposicionales elaborada por Ryle afronte con éxito la problemática exigencia de consumar sus análisis sin vaguedad ni circularidad. El cariz irremediablemente holístico de los estados mentales, o al menos de los conceptos que empleamos para su atribución –el segundo atolladero al que el conductismo lógico se habría visto abocado según el análisis de García-Carpintero (1995, *supra*)–, parecía hacer esa incompletud inherente al esquema de interpretación del vocabulario psicológico elaborado por Ryle, y conducía acaso por un camino aun más derecho hasta el funcionalismo. No es difícil apreciar que el hecho de que los conceptos con que atribuimos estados mentales se enmadejen sin cesar unos con otros hace impractible el análisis disposicional. El planteamiento del problema se retrotrae al menos hasta Chisholm (1957) y Geach (1957), aunque –si Block (1980a: 32) está en lo cierto– sería de nuevo en el decisivo trabajo de Putnam (1967a) donde cobraría toda su fuerza. La cuestión es que, tal como apunta Rodríguez (2001a) al hilo de una cita de Bechtel (1988: 125):

> […] el análisis de un término mental en términos disposicionales llega por lo general, más tarde o más temprano, a un nuevo término mental: estamos atrapados en un círculo de términos mentales. (Rodríguez 2001a: 87)

Quizá más taxativa es la formulación del problema que nos daba García-Carpintero (1995): si nos detenemos a detallar las disposiciones conductuales en que, dando por bueno el análisis ryleano, consistiría un estado mental cualquiera y tratamos de incorporarles referencias expresas a las circunstancias en que cada una de esas disposiciones se manifestaría –es decir, de precisar de forma exacta y exhaustiva las cláusulas condicionales, *caeteris paribus* incluidas, de los enunciados disposicionales

en cuestión, cosa que, como ya señalara Hampshire (1950), nadie parecía dispuesto a tomarse el trabajo de hacer–, lo que nos encontraremos es que:

> […] no hay ninguna posibilidad de enunciar las condiciones en que se darían los comportamientos pertinentes al caso sin hacer mención en ellas de otros estados mentales. (García-Carpintero 1995: 52)

Entre la afirmación de que la recurrencia de conceptos psicológicos en el análisis disposicional se dé sólo "por lo general" y la de que no haya "ninguna posibilidad" de que no se dé, media atinadamente el análisis originario de Chisholm (1957): la excepción vendría dada, acaso, por ciertos estados mentales que podríamos clasificar como *intenciones de ejecutar acciones básicas* –básicas, en el sentido de que no se realizan por mediación de ninguna otra. Así, por ejemplo, la intención de abatir los párpados (Rey 1997: 163), o la intención de abrir la boca, serían estados mentales para los que vale sin matices el análisis ryleano. Tan pronto como el contenido de la intención se despega del exiguo territorio de las acciones básicas –tan pronto como hablamos, por ejemplo, de la intención de coger un tren o de aprobar un examen–, su análisis disposicional comienza a plagarse de estados mentales concurrentes.

Desde luego, cuando no son intenciones sino creencias o deseos lo que intentamos releer bajo la óptica aconsejada por Ryle, es difícil incluso que la traducción de una atribución de actitud proposicional a enunciados sobre disposiciones dé sus primeros pasos. Así, para rendir cuenta de la creencia de que va a llover que –supongamos, siguiendo el ejemplo de Searle (1992: 34)– alberga un sujeto cualquiera, pronto nos veremos forzados a mencionar salvedades como que el sujeto, si se encuentra en casa, tenderá a cerrar las ventanas sólo si es que *desea* que la lluvia no empape la habitación, si es que *cree* que las ventanas abiertas dejan pasar la lluvia y que las cerradas no, o si es que *cree* estar en una casa con ventanas –entre otras muchas cosas que pueden llegar a ser tan tontas como incuestionables. Con trabas semejantes toparemos, naturalmente, a la hora de agregar al análisis la disposición a, si va a salir de casa, llevarse un paraguas, que sólo valdrá si el sujeto *desea* no mojarse, si cree que los paraguas sirven para evitar mojarse cuando llueve, si *cree* poseer un paraguas, etc., o incluso con la tendencia a, si se le pregunta al respecto, afirmar que va a llover, la cual sólo estará en vigor mientras el sujeto *desee* decir la verdad, *entienda* la pregunta, etc. El análisis de la creencia de que va a llover es, de hecho, el germen de una de las más tempranas formulaciones del problema, que surge cuando Geach (1957) discute a Ryle que la expectativa de lluvia pueda entenderse como una y la misma cosa que los comportamientos por él enumerados al relatar cómo "[…] the gardener who […] expects rain […] leaves the watering-can in the toolshed, keeps his coat handy, beds out more seedlings, and so on" (Ryle 1949: 175), aun cuando bien puede ser cierto que el jardinero, en efecto, haga todas esas cosas si, o porque, espera que llueva. El escollo insalvable en la propuesta de Ryle es, según Geach, que la descripción de la conducta que Ryle adopta da por sentado que conocemos las demás creencias y deseos del sujeto que puedan ser relevantes al caso.

En el fondo, Ryle está describiendo los comportamientos del jardinero como "acting as if you held such-and-such a belief", porque:

> In Ryle's example, this information is smuggled in by speaking of a *gardener*'s rain-expecting behavior (and tacitly assuming that the gardener is not e.g. a discontented or corrupted servant who wants the garden to be ruined). (Geach 1957: 8)

El vínculo entre este *problema del retorno de lo mental* y el carácter holístico de los conceptos de los que nos valemos para atribuimos estados mentales es patente: *porque* nuestras creencias y deseos están enmarañados como lo están, el análisis disposicional de unos remite irremisiblemente a otros, o, si se prefiere, la maraña de nuestras creencias y deseos y la inviabilidad del análisis disposicional son facetas de la misma cosa. De lo que se trata, a fin de cuentas, es de que –como enfatizara Rey (1997: 154)–, "[…t]ypically, *any particular mental state causes a particular behavior only in conjunction with (an often large number of) other mental states*".

Bajo esta luz, ensayos más ambiciosos –como el de Malcolm (1959) de analizar el concepto de tener un sueño como la propensión a relatar hechos ficticios al despertar– parecen irremediablemente malogrados. Tan notoria como sus vínculos con el carácter holístico de nuestros conceptos psicológicos es, en efecto, la gravedad del problema en relación con los intereses teóricos del conductismo lógico, que García-Carpintero (1995) consigna con nitidez:

> El conductista cree que los conceptos mentales son conceptos de disposiciones a conducirse de ciertos modos en ciertas circunstancias. Si esto es así, los estados mentales deben ser definibles sin remanente alguno en términos de tales disposiciones. Pero esta condición no puede ser satisfecha si las condiciones incluyen referencias a otros estados mentales. (García-Carpintero 1995: 53)

En la misma línea, Block (1980a: 32) formula el problema del retorno de lo mental como el de la inviabilidad de dotar de contenido a la cláusula *caeteris paribus* que acompañaría, explícita o implícitamente, a las traducciones disposicionales de atribuciones de estados mentales, sin incluir términos que se refieran a estados mentales.

> […B]ehaviorists […] attempted to define a mental state in terms of what behaviors would tend to be emitted in the presence of specified stimuli. E.g., the desire for an ice-cream cone might be identified with a set of dispositions, including the disposition to reach out and grasp an ice-cream cone if one is proffered, other things being equal. But, as functionalist critics have emphasized, the phrase "other things being equal" is behavioristically illicit, because it can only be filled in with references to *other mental states* […] (Block 1980a: 32)

Acaso más corrosivo aún es el destilado de las dificultades –o lo que "[…] parecen ser de hecho algo más que dificultades"– del conductismo lógico que había elaborado Putnam (1967a: 229), donde la cuestión no es ya que el análisis

disposicional de un estado mental *M* acabe tan ineluctable como furtivamente por apelar a otros estados mentales *N*, *O*…, sino al propio estado mental *M*. De lo que se trata es en definitiva –piensa Putnam– de la dificultad de:

> […] specifying the required behavior disposition [for the state of being in pain] except as "the disposition of X to behave as if X were in pain […]" (Putnam 1967a: 229)

Cuanto más convincente parece resultar la tesis de que el problema del retorno de lo mental devasta las perspectivas explicativas del conductismo lógico, más intrigante se torna, al menos a primera vista, la analogía que recalca Rey (1997: 163) entre el argumento del retorno de lo mental y la tesis de indeterminación de la traducción de Quine, un pensador de indiscutibles propensiones conductistas. En realidad, el propio Quine (1960: 279) señala que la tesis de indeterminación de la traducción –en este contexto, la tesis de que no existe evidencia conductual alguna que pueda permitirnos decidir entre un número indeterminado de atribuciones de actitud proposicional[73]– es inseparable del argumento de Chisholm según el cual "[…] no es posible abandonar el vocabulario intencional por el procedimiento de explicar sus miembros en otro lenguaje". Entender que Quine no está con esto labrando su propia derrota exige seguirle en uno de los requiebros más audaces de su pensamiento: el que le lleva a esgrimir las tesis de Brentano acerca de la naturaleza de la intencionalidad, *a contrario*, como "prueba de la falta de base de los giros intencionales y la vaciedad de una ciencia de la intención" (Quine 1960: 280, *supra*). El problema del círculo de lo mental no aparece, pues, a ojos de Quine sino como "[…] una tesis de Brentano, luminosamente desarrollada por Chisholm, […] que resulta directamente relevante para nuestras nacientes dudas acerca de las actitudes proposicionales y de otras locuciones intencionales" (Quine 1960: 279). Que las atribuciones coloquiales de creencias y deseos no se sometan al análisis disposicional, como demuestra Chisholm, es para Quine señal de la inconsistencia de dichas atribuciones. El conductismo que Quine habrá de proponer difiere, en suma, del de Ryle (1949) en su renuncia a la reducción del discurso psicológico ordinario a una teoría de la conducta adecuadamente reglada, lo que, por supuesto, precipita a Quine hacia una lectura eliminacionista del análisis disposicional. Es comprensible, entonces, que Quine –por lo demás, un holista convencido– tome partido por Chisholm en su escaramuza con Ryle, si bien dispuesto a extraer de la refriega una lección bien distinta de la que Chisholm auspiciaría[74].

---

[73] Con el trillado ejemplo del propio Quine: no existe evidencia conductual alguna que nos permita decidir entre la afirmación de que una persona que, cuando se encuentra en presencia de conejos, pronuncia en un idioma que nos es del todo desconocido la palabra "Gavagai", está pensando en un conejo, o la de que está pensando en una instancia de cunicularеidad, o en un estadío temporal de un conejo, o en el conjunto integrado de las partes de un conejo, etc.

[74] Es de rigor apuntar, sin embargo, que el pensamiento de Quine habría de apartarse por completo de estos planteamientos cuando, en 1990, acabara concediendo que:

En la estela de Quine, también Rey (1997) hace remontarse la cuestión de la pertinacia de los conceptos psicológicos en el análisis disposicional, al que parecen mostrarse refractarios, hasta los influyentes argumentos de Chisholm (1957). Pero sobrevolando intrépidamente las lindes entre conductismo lógico y conductismo psicológico, Rey ilustra el problema no con un ejemplo de análisis disposicional ryleano construido *ad hoc*, sino con una definición conductual de la noción de *expectativa* propuesta por Tolman de cara al "[…] análisis experimental y teórico de los determinantes de la conducta de las ratas en los puntos de decisión de un laberinto" que –según el ambicioso juicio del propio Tolman (1938: 34)– permitiría investigar ni más ni menos que, siquiera "[…] en esencia, […] todo cuanto es importante en psicología". Sin embargo –anota Rey–,

> […i]n Tolman's […] proposed definition of "a rat expects food at location L", defining a rat's *expectation* that there's food at L in terms of his moving towards it, works only if the rat wants food; and the rat's *wanting* food can be defined in terms of its moving towards L only if it *expects* there's food at L. Insofar as this is true, the prospects of a definition of a *single* informational or *single* directional state in terms of behavior seem dim. (Rey 1997: 154)

Como se ha visto, buena parte de la censura del conductismo que impulsó los planteamientos cognitivistas descansaba sobre el convencimiento de que cómo sea el ambiente en que se desenvuelve el organismo es a menudo menos relevante de cara a la explicación de su conducta que cómo el organismo lo represente para sí[75]. Pues bien, la traslación de la controversia consumada por Rey hace patente que las críticas de Chisholm (1957) o Geach (1957) al análisis disposicional del vocabulario mentalista bien pueden verse como la formulación más general y temprana de la misma convicción que impulsara uno de los argumentos cardinales de Chomsky (1959) contra Skinner (1957): a saber, que las estrategias explicativas empleadas por Skinner no prescindían –como él pretendía– de referencias a estados internos del sujeto, sino que enmascaraban dichas referencias bajo imprecisos conceptos de clases de estímulo o de respuesta –tal como sucede, si Rey (1997) está en lo cierto, en el trabajo de Tolman (1938) en relación con el concepto de *expectativa*. En efecto, aunque las críticas de Chomsky suelen reconocerse con justicia como decisivas en la desacreditación del conductismo, es de rigor advertir que su argumento en este

---

> Brentano was right about the irreducibility of intentional discourse. […] It implements vital communication and harbors indispensable lore about human activity and motivation, past and expected. Its irreducibility is all the more reason for treasuring it: we have no substitute. (Quine 1990: 71)

Poco antes, en "States of Mind", Quine (1985: 6) había renunciado a trazar una distinción clara entre un materialismo de corte reduccionista y uno eliminacionista, aseverando que no alcanzaba a ver difeencias entre ambos. Con las dificultades del reduccionista para guardar el precario equilibrio que le impide sucumbir al eliminacionismo nos encontraremos reiteradamente *infra*.

[75] Ése fue sin duda uno de los motivos de la acentuada orientación internista que, según se verá con detalle *infra*, acompañaba a las primeras versiones del funcionalismo cognitivista.

punto parece una estricta aplicación, contra la teoría de cómo aprendemos el lenguaje ideada por Skinner, del razonamiento que habían dejado perfilado Chisholm y Geach, según el cual la palabrería psicológica denostada por Ryle (1949) resurgía tan pronto como se intentaba dar razón de las cláusulas *caeteris paribus* insondablemente laxas que, de forma implícita, acotaban sus análisis disposicionales. De manera que en torno a la tesis del retorno de lo mental confluyen tanto el convencimiento de los psicólogos cognitivos respecto a la importancia de la representación de las características del entorno por parte del sujeto a la hora de explicar su conducta, como las críticas de Chomsky en cuanto a la propensión de los conductistas –y en particular de Skinner (1957)– a traficar con dichas representaciones bajo el rubro de vaguísimas descripciones de estímulos y respuestas.

Bien puede apreciarse esta constelación conceptual en el ejemplo que abre las argumentaciones de Pylyshyn (1984): un peatón contempla como un automóvil se estrella contra un poste, vacila, mira dentro del coche, se apresura en llegar a una cabina telefónica cercana y marca los dos primeros dígitos del número de emergencias; parece obvio que su siguiente conducta será marcar el tercer dígito. Ahora bien, esta obviedad descansa sobre un buen número de presupuestos que orbitan en torno a la descripción del acontecimiento, muchos de ellos con implicaciones mentalistas; de ahí que, de acuerdo con Pylyshysn:

> […] no account stated solely in behavioral terms (that is, in terms that do not incorporate the person's knowledge or goals) can make such a prediction. The reason is, if a systematic account is to connect the prediction […] with the "stimulus" conditions, such an account must at the very least mention: that the pedestrian *interpreted the scene as an accident*; that the pedestrian *knows* or *remembers* the phone number; and that the pedestrian's behavior is an instance of the category "phone for help". (Pylyshyn 1984: 7)

Salvo porque Ryle apuntaba a una redefinición conceptual de la noción de expectativa y el teórico conductista que sirve de contendiente a Pylyshyn a la predicción de un fragmento de conducta –por lo demás, justo la divergencia esperable dadas las diferencias entre el conductismo lógico y el psicológico, sea de índole metodológica o "radical"–, la estructura del ejemplo confeccionado por Pylyshyn es exactamente análoga a la del caso del jardinero que irrigara la discusión entre Ryle (1949) y Geach (1957). Incluso la exhibición de contrafácticos más o menos pintorescos recuerda estilísticamente a Geach y su jardinero ruin: la predicción de la conducta del peatón quedaría en entredicho, por ejemplo, si contáramos que éste sabe que lo que ha contemplado es parte de un rodaje televisivo.

Como el análisis de Geach (1957) hizo ya patente adelantándose al de Chomsky (1959), los caminos por los que los matices psicológicos pueden infectar la descripción de estímulos y respuestas son, poco menos que inescrutables. La afinidad del razonamiento de Pylyshyn con las imputaciones que Chomsky (1959) levantara contra Skinner se perfila también de inmediato. De hecho, tales imputaciones cobran aquí una de sus formas más incisivas: los conceptos básicos del conductismo, como estímulo o respuesta, son *en la práctica* conceptos cognitivistas.

> In practice –though not in theory– the way behaviorists get around the need for such a mentalistic vocabulary is to load the description of conditions or behavior with cognitive content. Instead of a physical vocabulary, they use a "behavioral" vocabulary in which things are described as *stimuli*, *responses*, *reinforcers*, and so on. The reason this works is that, in practice, these categories are cognitive: What serves as the functional stimulus depends on how a person interprets the situation […]. Similarly, what constitutes the response is implicitly cognitive. Some particular bit of movement […] does not count as a "response," only movements intended a certain way are counted. (Pylyshyn 1984: 8-9)

Por último, el énfasis en que cómo el organismo se represente su entorno es mejor predictor de su conducta que cómo resulte ser dicho entorno, aunque salta a la vista en el planteamiento del ejemplo, es explicitado por el propio Pylyshyn:

> The critical aspect of the connection between stimulus conditions and subsequent behavior –the aspect that must be incorporated in the theoretical account if the latter is to capture the systematicity of the person's behavior– is that it is (a) the environment or the antecedent event *as seen or interpreted by the subject*, rather than as described by physics, that is the systematic determiner of actions; and (b) actions performed with certain *intentions*, rather than behaviors as described by an objective natural science such as physics, that enter into behavioral regularities. (Pylyshyn 1984: 9)

Desde luego, no es difícil adivinar también en el pensamiento de Pylyshyn la proclividad a una concepción internista de la semántica de los estados mentales que deriva de este énfasis en la primacía de la representación. Más transparente es, con todo, la convicción de Pylyshyn de que éstas eran en esencia las razones de la inviabilidad del conductismo: "It is this remarkable degree of *stimulus-independent control* of behavior" –diría, con términos intensamente reminiscentes de Miller, Galanter y Pribram (1960)– "that has been the Achilles' heel of behaviorism" (Pylyshyn 1984: 10).]

El caso es que cabe entender el desarrollo del funcionalismo como el camino marcado por la asunción de que el problema del retorno de lo mental desarbolaba no sólo el conductismo lógico, sino cualquier concepción de nuestros estados psicológicos que no contemplara sus relaciones recíprocas. Así parece entenderlo, entre otros, Fodor (1981a: 9, *infra*) cuando destaca el carácter relacional de lo mental como principal lección que el funcionalismo habría extraído de los análisis conductistas.

Es discutible, con todo, que explicitar referencias circulares que estaban veladas en una explicación valga por dejar la explicación expurgada de dichas referencias. Si el funcionalismo reclama haber vencido la circularidad que corrompía la explicación disposicional de los estados mentales, ha de contar con algún argumento que respalde la idea de que los elementos que cerraban el círculo pueden quedar neutralizados incorporándolos abiertamente a la explicación. Una circularidad explícita –cabría conceder– será a determinados efectos más benigna que una implícita, pero deshacer la circularidad sin duda exigirá algo más que develarla

y declararla extinta por *fiat*: el cognitivismo debe arrostrar, pues, el retorno del problema del retorno de lo mental.

Además, sobre el modelo de explicación psicológica basado en la atribución al organismo de estados internos que promovía el funcionalismo pesaban precisamente severas acusaciones de circularidad. Los propios conductistas –Ryle (1949), Skinner (1953)– habían blandido una y otra vez contra la explicación mentalista diversas versiones del argumento según el cual ésta sería, en cualquiera de sus formas, una variante particularmente tosca de explicación circular: dar cuenta de la conducta agresiva merced a la *irritabilidad* del sujeto –la cual se deduce precisamente de la frecuencia con que tienen lugar tales conductas agresivas–, o describir la *inteligencia* como causa de las conductas que nos permiten detectarla –es decir, de conductas inteligentes– sería, para Ryle o Skinner, tan clarificador como la apelación a la *virtus dormitiva* del doctor de *Le Malade Imaginaire*. Todo ello, desde luego, acrecienta la exigencia de que el funcionalismo muestre cómo puede esquivar la circularidad que por uno u otro derrotero se le imputa.

Pues bien, si la circularidad diagnosticada por Ryle y Skinner afecta también, como argumentaron Chisholm (1957) y Geach (1957), al propio planteamiento conductista –que no hace sino enmascararla–, el tránsito de una circularidad develada, ya explícita, a una circularidad extinta se gesta, de acuerdo con el planteamiento funcionalista, en el seno de la teoría de autómatas y de las investigaciones de Alan M. Turing, donde se nutre de la noción de *definición simultánea*. La idea, groseramente perfilada por el momento, viene a ser que la dañina circularidad que aqueja a una teoría que apela a los mismos fenómenos que pretende explicar quedaría abolida cuando dichos fenómenos vienen *simultáneamente* explicados por la propia teoría: ésta, entonces, se sostendría a sí misma como una prodigiosa arquitectura exenta, reflejando no ya la pobreza de nuestros conocimientos sino, al contrario, el intrincadísimo tejido de la naturaleza. Que las teorías psicológicas pudieran construirse así es, como se verá, una faceta de las intuiciones con las que Putnam inauguraría la concepción funcionalista de lo mental.

Una táctica muy parecida es la que propugna Rey (1997), en el contexto de su esforzada defensa de un realismo intencional de raigambre fodoriana frente a los embates del instrumentalismo y el eliminacionismo. Tras recordar la "dificultad técnica" (Rey 1997: 154) que Chisholm (1957) habría desenterrado de los cimientos explicativos del conductismo lógico –ya sabemos: la irremediable reaparición de conceptos psicológicos al desplegar el inventario de disposiciones presuntamente equivalente a un estado mental cualquiera–, Rey procura mitigar los daños:

> These particular observations of Chisholm's, however, aren't quite as devastating to A[nalytical] B[ehavior]ism as they might initially appear. For it turns out to be technically open to the A[nalytical] B[ehavior]ist to propose defining *all mental terms simultaneously* […]. (Rey 1997: 155)

A primera vista, que el conductismo lógico parezca poder reponerse de la objeción planteada por Chisholm con una argucia de esta índole da la razón a Searle (1992: 34), que relega el problema del retorno de lo mental al ámbito meramente técnico. Esto no significa, sin embargo, que Searle –como insinúa Rey (1997: 163)– pretenda desproveerlo de consecuencias de calado. Antes bien, lo que Searle (1992) pretende es contraponer las múltiples objeciones técnicas alegadas contra el conductismo a la objeción "de sentido común" según la cual el conductismo "[…] niega la existencia de estados mentales internos aparte de la conducta externa", lo cual, a su entender, sería evidentemente absurdo, en tanto que habla "[…] contra la experiencia ordinaria de cómo es ser un ser humano" (Searle 1992: 35). La objeción de sentido común, según aprecia Searle, es la más embarazosa para el conductismo –la que hace al conductista blanco de burlas y chistes–, y su aparición habría sido mucho más temprana. Searle, de hecho, la retrotrae hasta el influyente trabajo de Ogden y Richards (1923) sobre la noción de significado, que sin duda leyera en el transcurso de sus primeras investigaciones en torno a la teoría de los actos de habla, y del que provendría también (Searle 1992: 250, *cf.* también Lycan 1987: 4) la caracterización burlesca de los conductistas, y de Watson en particular, como pensadores forzados por la lógica de sus propias teorías a "[…] postular una ausencia general de sensibilidad" (Ogden y Richards, 1923: 47)[76]. Pero de nada de esto se sigue que las "objeciones técnicas" sean inocuas, o que tengan sólo consecuencias técnicas[77].

En cualquier caso, tal como el propio Rey apunta a renglón seguido, la ejecución de la maniobra técnica que lo rescata del círculo descrito por Chisholm encamina de todas todas al conductismo lógico hacia un planteamiento de rotundo aire funcionalista: esa idea de definición simultánea es –decíamos– una de las claves del análisis de lo mental enarbolado por el funcionalismo. Dicho de otro modo: la soltura con que el conductismo lógico parece zafarse de la crítica de Chisholm no es señal de que ésta lo deje ileso sino, más bien, de que lo aboca a ceder ante el funcionalismo.

---

[76] Como salta a la vista, "[…] postular una ausencia general de sensibilidad" es un giro que deja escapar buena parte de la sorna de la expresión que Searle recuerda haber encontrado en el trabajo de Ogden y Richards, quienes acusaban a Watson, literalmente, de *fingir anestesia general* ("[…] feigning general anesthesia", reza la edición inglesa de 1949 en su página 23). En Lycan (1987: 4) se añade otro punto sarcástico –ese fingimiento sólo resulta obligado para los conductistas *cuando están de servicio* ("[…] when on duty")–, que además se acompaña de la observación de que "[…m]any people understood Behaviorism as being a doctrine that no one could seriously believe".

Un tono muy parecido, pero velando una burla quizá aun más severa, puede apreciarse cuando Popper y Eccles (1984: 106) apuntan que "[…] Ryle seems to have tried to observe himself, but apparently he did not succeed […]".

[77] Sea como sea, resulta curioso que Searle (1992: 35) tome nota de los razonamientos modales de Putnam (1963a) contra el conductismo lógico como expresión más o menos sofisticada de esta objeción de sentido común, cuando cabe interpretarla también –como veíamos en García-Carpintero (1995: 53, *supra*)– como fruto de las críticas, de orden innegablemente técnico, respecto a la confusión entre efectos de lo mental y constituyentes de lo mental de la que sería presa el conductismo lógico en su análisis de la noción de disposición conductual.

En efecto, la artimañana de proponer una definición simultánea de todos los conceptos referidos a estados mentales, de la que según Rey (1997: 163) podría valerse el conductismo lógico para esquivar la acometida de Chisholm, aparece en la feroz mirada retrospectiva de Putnam (1997) sobre sus propios trabajos fundacionales como una de las claves de arco del funcionalismo:

> A formalism for computation theory *implicitly defines* each and every computational state by the totality of its computational relations (e.g., relations of succession, or probabilistic succession) to all other states of the given system. In other words, the whole set of computational states of a given system are *simultaneously implicity defined*; and the implicit definition *individuates* each of the states, in the sense of distinguishing it from all other computational states. (Putnam 1997: 35)

Más exactamente –piensa Putnam–, sería una vaga idea de definición simultánea lo que pronto habría de imponerse sobre la "idea original del funcionalismo", que era a su juicio la de que:

> […] our mental states could be identified with computational states, where the notion of a computational state had already been made precise by the preexisting formalisms for computation theory, for example, the Turing formalism or the theory of automata. (Putnam 1997: 35)

Pero las insuficiencias de dichos formalismos para dar cuenta de un buen número de peculiaridades de los estados mentales –desveladas, entre otros, por Block y Fodor (1972a, *infra*)– nos habrían forzado a reemplazarlos por la inverosímil doctrina de un formalismo normal de descripción psicológica, una "teoría psicológica ideal" en virtud de la cual todo estado mental posible quedaría sujeto a todos los demás por medio de la ceñida malla de la definición simultánea. Que tal cosa era fatalmente inverosímil resultaría claro para Putnam cuando contemplara la trayectoria del funcionalismo treinta años después de sus decisivos escritos de 1967[78], puesto que:

> […] no psychological theory individuates or "implicitly defines" its states in this sense. […] No actual psychological theory has ever pretended to provide a set of laws that distinguish, say, the state of being jealous of Desdemona's fancied regard for Cassio from every other actual or possible propositional attitude. (Putnam 1997: 35)

[78] Las reflexiones de Putnam bien podrían haberse producido a la luz de las críticas contra el "materialismo promisorio" vertidas por Popper y Eccles (1984: 96-97). Precisamente contra la concepción popperiana de la explicación histórica, había esbozado von Wright (1971: 179) objeciones parecidas: ante la insinuación de Popper (1945) de que leyes aparentemente triviales de, digamos, la sociología militar –como que la superioridad numérica asegura *caeteris paribus* la victoria– intervienen en la explicación de hechos históricos –en el ejemplo, la división de Polonia en 1772–, von Wright señala irónico con qué destreza "[…] defenders of the covering law theory of historical explanation" –Popper era uno de los más destacados– "succeed in evading relevant examples". Resulta claro que su preocupación es, en el fondo, idéntica a la de Putnam (1997, *infra*), que en una grácil pirueta histórica, parece inspirada en Popper y Eccles (1984).

Pero el veredicto no es del todo exacto. Acaso con la vehemencia propia de un trabajo de afán netamente retórico, Watson había dejado escrito en 1913 que:

> […] en un sistema de psicología completamente elaborado, dada la respuesta pueden predecirse los estímulos; dados los estímulos, puede ser predicha la respuesta. (Watson 1913: 167)

Por supuesto, Watson no aspiraba a diferenciar un particularísimo sentimiento de celos de entre la totalidad de estados intencionales posibles o efectivos, pero sí –o eso parece– a poder predecir, de acuerdo con determinadas leyes, una particularísima conducta de celos partiendo de una de entre la totalidad de configuraciones estimulares efectivas o posibles, o viceversa. Es obvio que la predicción de que el sujeto emitirá tal conjunto de conductas si se encuentra ante tal conjunto de estímulos presupone la identificación de dichas conductas como conductas del tipo pertinente al estímulo –conductas celotípicas respecto de tal otro sujeto, en el ejemplo–, o viceversa; es decir, su diferenciación de cualquier otro conjunto de conductas –o estímulos, según el caso– posibles o efectivas. Salvada la distancia, así pues, entre la visión conductista y la cognitivista del objeto de la psicología, parece claro que las ambiciones explicativas de una y otra escuela son parejas. Desde luego, las ambiciones depositadas a veces en la neurofisiología no se quedan atrás: bien conocido es, por citar un ejemplo seminal, el augurio de Feigl (1958: 442) según el cual el estudio del cerebro debería en el futuro proveernos de "[…] complete deductive derivations of the behaviour symptoms of various central states" –la revelación de las "maravillosas complejidades" de nuestra vida interior a la luz de una "cinemática y una dinámica materialista" finalmente maduras, que vaticina Churchland (1984/1988: 256, *supra*) sería fruto sin duda de avances de parecida envergadura.

Aunque Putnam (1993), entonces, haya exagerado el carácter idiosincrásico de la arrogancia cognitivista en su pretensión de omnisciencia, parece razonable dar por bueno que el funcionalismo nace –en uno de sus manantiales al menos– bajo la forma de la asunción de que definir la noción de estado mental sin verse forzado a incluir en el *definiens* otros estados mentales es, contra el conductismo, un logro inalcanzable: la asunción, pues, de la circularidad inherente de lo mental. Así que, si el examen de Putnam es atinado, la espada de Chisholm y Geach, que había decapitado al conductismo lógico, acabaría siendo también fatídica, después de todo, para una concepción de la mente, la funcionalista, que cobró forma precisamente para tratar de superar las taras de la idea de lo mental que venían propugnando los conductistas.

En la herencia conductista de la teoría computacional de la mente, si todo esto es así, habría residido el mal congénito que a la larga acabaría con ella. Pero lo que en todo caso queda acreditado –aun si Putnam se hubiera mostrado demasiado severo

en su juicio– es que las raíces del funcionalismo ahondan en las audaces investigaciones de Ryle, por mucho que éstas aparecieran al principio –lo hemos visto, por ejemplo, en Fodor (1968)– como heraldos de un enemigo acaso no tan porfiado[79].

Despliegue y alcances del fisicalismo

Descuidadamente, se describe a veces el auge del cognitivismo en psicología como un resurgimiento del interés por el funcionamiento del cerebro, al que los años de hierro del conductismo habrían relegado al mediocre papel de una suerte de estación de relés, de una intricada carta de reflejos a la que, al igual que a una mente a la que esporádicamente se daba por inexistente, debía por mor del método considerarse simple y llanamente *caja negra*. La psicología cognitiva habría llegado entonces, de la mano del avance imparable de nuestros conocimientos en torno a la anatomía y la fisiología del sistema nervioso –que, por lo demás, era ya una avalancha desde tiempos de du Bois-Reymond– a redimirnos de la doble ignorancia de una misma realidad, la de los procesos mentales y los cerebrales, que la doctrina conductista imponía.

Por mucho que insistiera –contra el tópico– en que el organismo "[…] no puede ser adecuadamente tratado como una caja negra" (Skinner 1976: 223), lo cierto es que Skinner relegaba sin titubear a los mecanismos y procesos internos del organismo al papel de meros transmisores en la cadena de causas y efectos que engranaba estímulos y respuestas:

> Unless there is a weak spot in our causal chain so that the second [neurological] link is not lawfully determined by the first [environmental stimuli], or the third [behavior] by the second, the first and third links must be lawfully related. […] Valid information about the second link may throw light on this relationship but can in no way alter it. […It is] external variables of which behavior is a function. (Skinner 1953: 35)

Por supuesto, la confianza de Skinner en que no hay eslabón débil en esta cadena de causas y efectos es un ejemplo de la ingenuidad que la insistencia de los pioneros del cognitivismo en la primacía de la representación del entorno como determinante de

[79] Tanto peor, dicho sea de paso, si tiene razón Smith (2002b: 242) en que la noción de contenido semántico, que el funcionalismo asume al postular estados mentales de naturaleza intencional o representacional, es anterior a nuestras categorías ontológicas comunes: entonces la circularidad de la teorización cognitivista será aun más severa. Al emplear términos teóricos para designar estados mentales que se refieren a objetos, eventos, hechos mundanos, etc., una psicología de inspiración funcionalista, como la cognitiva, estaría reposando sobre nociones inexplicadas de objeto, evento, hecho, etc., siendo así que aclarar esas nociones requerirá mención de estados intencionales. Con las palabras del propio Smith (2002b: 242): "Because ontological categories are in part intentionally constituted, attempting to explain representation while dining out on ontology is […] fatally circular". Razonamientos parecidos al de Smith, pero al hilo de la noción de causalidad, se ensayarán repetidamente *infra*.

la conducta venía a desafiar, a menudo desde el propio seno del conductismo –ingenuidad, claro está, sólo a ojos de los cognitivistas. El eslabón débil que Skinner ignora residiría precisamente en que la conducta no es una función de las variables externas, sino de su representación interna, y tanto entre la variable externa y su representación (*cf*. Fodor 1968: 55, *supra*) como entre la representación y la conducta (*cf.* Lashley 1951: 230 y Guthrie 1953: 172 *apud* Miller, Galanter y Pribram 1960: 21 y 19 respectivamente, *supra*) pueden concurrir multitud de procesos que hagan impracticable el atajo ensayado por Skinner: esto es, *procesos psicológicos*.

Con todo, incluso un estudio somero de la historia de las muchas vertientes del conductismo hace patente que su caracterización, con ánimo de generalidad, bajo el despectivo rubro de la caja negra es de todo punto injusta. No mucho más fiel a los hechos resulta, en cualquier caso, la idea de que cognitivismo y fisicalismo llegaran de la mano para corregir, en el mismo sentido, el rumbo de la investigación en psicología y de la reflexión en filosofía de la mente –como hace, por ejemplo, Clapin (2002a) cuando, con aire de recapitulación, anota que "han pasado ya cerca de cincuenta años […]

> […] since the cognitive revolution in the scientific study of the mind (marked by the rise of artificial intelligence and cognitive psychology), and the (re)turn to materialism and physicalism in the philosophy of mind, as marked by Smart's (1959) and Place's (1956) statements of mind-brain identity theory. (Clapin 2002a: 2-3)

Ni siquiera sería del todo exacto retratarlos como dos desplazamientos simultáneos de nuestra concepción de lo mental que, aunque apuntaran en direcciones distintas, acabarían por confluir. Más verosímil resultaría, probablemente, situar uno de las manantiales del pensamiento funcionalista, que habría de confluir con la naciente psicología cognitiva, no en la formulación primitiva de la tesis de identidad psicofísica, cuya compatibilidad con el funcionalismo cognitivista sigue siendo objeto de controversia, sino precisamente en las renuentes restricciones y matices al alcance de dicha tesis que comienzan a entreverse en las posiciones de Smart (1959) o Armstrong (1966, 1968). Dichas restricciones cristalizarían en la noción de identidad de rol causal a la que David K. Lewis (1966, 1972), aun sin desprenderse nunca de su inquebrantable confianza en la lectura más ambiciosa de la tesis de identidad psicofísica, habría de dar preeminencia –convirtiendo así la distinción entre rol y ocupante de un rol, en cierto grado a su pesar, en uno de los pilares de la incipiente concepción funcionalista de lo mental–, pero también en la crítica frontal de Putnam (1967a) a esa interpretación fisicalista de la tesis de identidad defendida por Lewis, crítica que daría carta de naturaleza a un funcionalismo entendido ya como alternativa al fisicalismo.

Mirémoslo con mayor detenimiento. El trayecto de retorno al materialismo que inaugura Place (1956) –retorno, entiéndase, desde las comarcas del conductismo lógico– brota de la dificultad que arrostraba Ryle a la hora de dar cuenta en términos disposicionales de la diferencia entre –pongamos por caso– la experiencia de

contemplar una extensión de azul cobalto y otra de azul de Prusia. Los fenómenos psicológicos que por un tiempo se dio en llamar "sensaciones brutas" –o "crudas": *raw feels*[80]–, como el dolor, se prestaban dócilmente a argumentos modales como los que compendiaría Putnam (1963a, *supra*), en los que se muestra que es tan concebible que todo el entramado de disposiciones conductuales ligado a un estado mental de esa índole esté presente sin que lo esté la experiencia subjetiva (los super-actores imaginados por Putnam), como que lo esté la experiencia subjetiva sin que se muestre ni una sola de las disposiciones esperables (los super-super-espartanos), o sólo las de naturaleza verbal (los super-espartanos). La razón, según pensaba Place, es que el carácter cualitativo de una sensación –su *quale*, se decía ya entonces, en la estela de Clarence I. Lewis (1929)– tiene una realidad instantánea, actual, que no se compadece bien con la dilación que es propia de las disposiciones: el punzante aguijón de un cierto dolor, o la profundidad de un matiz u otro de azul, mal podrían consistir en la tendencia a hacer tal o cual cosa en un futuro hipotético, puesto que parecen estar presentes plenamente en el momento mismo en que los sentimos.

Uno de los mayores logros de la geografía conceptual de lo mental desplegada por Ryle –acaso el mayor– fue, según la valoración retrospectiva de Place (1999), que la amplitud de su análisis le permitiría esquivar la conclusión de que ningún lenguaje podría referirse a eventos estrictamente privados, que tentó a Wittgenstein (1953). Así pues, Ryle *habría podido* conceder que existen, además de verbos psicológicos que denotan disposiciones comportamentales –casi todos– y de otros que denotan logros epistémicos o de otra índole (*achievement words*, Ryle 1949: 149), aun otros que denotan actividades duraderas, como la atención (*heed*, Ryle 1949: 135-149) o eventos mentales de carácter –digamos– íntegramente episódico (*a clockable occurrence*, Ryle 1949: 140; cosas que, como los sentimientos, "[…] vienen y van o crecen y se desvanecen en pocos segundos", Ryle 1949: 100):

> […] a small minority of [psychological] terms that refer or contain a reference to an event or process taking place beneath the individual's skin to which he or she has some kind of "privileged access" […] (Place 1999: 380)

Pero debido al mismo prurito conductista en aras del cual Watson o Skinner (*supra*) habrían errado en su intento de remendar subrepticiamente en sus discursos teóricos el reconocimiento de tales eventos internos bajo la forma de epifenómenos, Ryle, "[…] obviously embarrassed by having to make the concession [above] […] does his best to minimize its magnitude and significance" (Place 1999: 380). En el caso de las sensaciones brutas, en particular, la expeditiva minuta rubricada por Ryle aboga por dejar de lado la cuestión: toda vez que en la descripción cotidiana de nuestras sensaciones nos valemos de un vocabulario referido al aspecto de lo percibido, no a

[80] La locución hace fortuna a partir de Feigl (1958), quien, como nos recuerda Güzeldere (1997: 56), podría haberla tomado de Edward C. Tolman (1932/1967: 250-251) : "Sensations, says the ortohodox mentalist […] are immediate mental givens, 'raw feels'. They are unique subjective suffusions of the mind" –escribe Tolman al describir la posición de sus adversarios dialécticos.

propiedades de la propia percepción, "[…] there is nothing 'mental' about sensations" –concluye audazmente Ryle (1949: 204)[81]. Desde esta modesta perspectiva, la aportación del propio Place (1956) se habría limitado a desdramatizar una concesión que estaba ya al alcance de la mano de Ryle, dándole carta de naturaleza en una concepción mecanicista del mundo al identificar esos estados internos que nos embargan *hic et nunc*, que conocemos personalísimamente pero que a duras penas podemos describir, con algo cuya naturaleza material resulta tan incontestable como es el caso de los estados del sistema nervioso.

De esta manera, si bien parecería haber encontrado más el barlovento del fisicalismo que el del funcionalismo, la cuestión de las sensaciones brutas habría de incorporarse al diagnóstico de las dolencias terminales del conductismo lógico. Al lado del hecho de que las actitudes proposicionales se nos presenten como el fundamento de ciertas disposiciones más que como las disposiciones mismas, y de que lo hagan en medio de una maraña a duras penas penetrable –las dos dificultades que García-Carpintero (1995: 53, *supra*) registraba como claves de la transición del conductismo lógico al funcionalismo–, debería consignarse entonces, con Rorty (1979: 101), el hecho de que

> […] many philosophers who agreed that Ryle had shown that beliefs and desires were not inner states agreed also that he had left raw feels untouched, and thus that a choice still had to be made between dualism and materialism.

Identificar las sensaciones crudas con estados físicos del cerebro permitiría –pensaba Place– salvaguardar el análisis disposicional de las creencias y los deseos detallado por Ryle, sin temor a desincrustarlo de una visión inequívocamente materialista de las relaciones entre la mente y el cerebro. Tendríamos, así, que:

> In the case of cognitive concepts like "knowing," "believing," "understanding," and "remembering," and volitional concepts like "wanting" and "intending," there can be little doubt, I think, that an analysis in terms of dispositions to behave (Wittgenstein, 1953; Ryle, 1949) is fundamentally sound. On the other hand, there would seem to be an intractable residue of concepts clustering around the notions of consciousness, experience, sensation, and mental imagery, where some sort of inner process story is unavoidable. (Place 1956: 44)

Cada propiedad consciente, ya fuera aquel punzante dolor o el matiz preciso del azul cobalto, que pudiera darse en un organismo resultaría coextensiva con alguna propiedad física de su sistema nervioso, y esas coextensividades, desentrañadas empíricamente, quedarían aglutinadas en las leyes-puente de un estricto programa de reducción teórica, a la manera de Kemeny y Oppenheim (1956) o Nagel (1961). En suma, según la formalización canónica recogida por Liz (1995: 223), lo que se venía a

[81] Incluso en el delicado ámbito de las sensaciones brutas habría quedado, entonces, prefigurada en Ryle la aproximación que a la larga se desplegaría en el seno de la tesis de identidad psicofísica: el análisis temáticamente neutral auspiciado por Smart (1959, *infra*).

proponer era que *(M)(∃F)(x)(MxFx)*: para toda propiedad mental *M* existe una propiedad física *F* (o, si se prefiere, para todo tipo de estado mental *M*, definido por la posesión de dicha propiedad, existe un tipo de estado físico *F*, definido por la posesión de la propiedad correspondiente) tal que para todo *x*, *x* es *M* (o pertenece a *M*) si y sólo si *x* es *F* (o pertenece a *F*). Lo que se postula, dicho de otro modo, es una identidad entre tipos de estados mentales –clases de equivalencia de estados mentales delimitadas según criterios psicológicos– y tipos de estados cerebrales –clases de equivalencia de estados cerebrales delimitadas según criterios neurológicos[82].

Ahora bien, la distinción entre un rol causal y aquello que lo desempeña, que vendría a convertirse en uno de las herramientas cardinales del pensamiento funcionalista, estaba ya casi madura en las primeras versiones de la tesis de identidad compuestas por Smart (1959): a su juicio, lo que la investigación empírica revelará es que los (tipos de) estados mentales que inicialmente caracterizamos, mediante el análisis conceptual del discurso ordinario, en términos de su *papel* como mediadores entre estímulos y respuestas, son de hecho (tipos de) estados del sistema nervioso central. Como recordaría Lycan (1987: 7), por ejemplo:

> […] if we ask what a pain is, we initially characterize it in terms of its typical external causes and effects: it is the sort of inner state that results from damage and issues in withdrawing-and-favoring behavior; when we open up an organism and look inside, we find what sort of state that is, e.g., a firing of c-fibers.

Además, adoptar una posición estrictamente fisicalista en cuanto atañe al carácter cualitativo de las sensaciones abría el camino para dotar a lo mental de la eficacia causal que la interpretación de Ryle le denegaba. Subyacía, como hemos visto, a la argumentación de Putnam (1963a) la convicción de que *lo que sentimos* no puede identificarse con el modo en que sentirlo nos hace comportarnos, porque es precisamente la causa de que nos comportemos de ese modo. Ésa era, no en vano,

---

[82] Para una definición formal de la noción de *clase de equivalencia*, *cf*. Mosterín (1984: 78-79): en suma, el conjunto de todos los miembros de un conjunto *A* que guardan con un miembro dado, *x*, una *relación de equivalencia R*, se denomina clase de equivalencia de *x* respecto a *R*, y determina una *partición* del conjunto *A*. Una relación de equivalencia en el conjunto *A* es una relación entre pares de miembros de *A* que es reflexiva (todo *x* la mantiene consigo mismo), simétrica (si *x* la mantiene con *y*, entonces *y* la mantiene con *x*, y viceversa) y transitiva (si *x* la mantiene con *y*, y viceversa, e *y* con *z*, y viceversa, entonces *x* la mantiene con *z*, y viceversa). Una partición de *A* es una familia de subconjuntos no vacíos de *A* tal que la unión de todos ellos es idéntica a *A* (es decir, un *recubrimiento* de *A*), y que carecen de elementos comunes entre sí (es decir, que además de no vacíos son disjuntos, *cf*. Mosterín 1984: 19). Para una definición más informal de la noción de *tipo*, *cf*. Mosterín (1984: 107): un tipo es un conjunto introducido mediante una condición necesaria y suficiente. En la discusión científica y filosófica acerca del alcance con que debe entenderse la tesis de identidad psicofísica, suele emplearse la noción de tipo incluso en circunstancias en que no cabe suponer que existan tales condiciones necesarias y suficientes –incluso, de hecho, en circunstancias en que es precisamente la existencia de dichas condiciones lo que está sometido a debate–, por lo que sería más adecuado hablar de clases de equivalencia. Mientras esta convención resulte inofensiva, podemos pasar por alto el matiz.

una de las fisuras en el armazón conceptual de Ryle (1949) en las que Geach (1957: 6, *supra*) había estado hurgando. Si las sensaciones brutas eran, sin embargo, ni más ni menos que estados neurofisiológicos, su legitimidad como causas quedaba fuera de toda sospecha. Al contrario, tal como sugiere Graham (2007), resultaría incluso "[…] tentador postular que […]"

> [...] the qualitative aspects of mentality affect non-qualitative elements of internal processing, and that they, for example, contribute to arousal, attention, and receptivity to associative conditioning. (Graham 2007: §5)

Si Place (1956) escindía el territorio de lo mental en dos vastas regiones –lo intencional, lo cualitativo– cuyas geografías debieran trazarse con distintos aparejos –el análisis disposicional, la identidad psicofísica–, Feigl había de argumentar poco después que ambos atlas –el conductista, el fisicalista– revelaban en realidad los mismos parajes. Merced a una medida restauración de la costumbre de Russell (1911) de distinguir entre conocimiento por familiaridad (*by acquaintance*) y por descripción (*by description*) –que se avenía bien con el decidido recurso de Feigl (1958, *infra*) a la distinción fregeana entre sentido y referencia para aclarar las tesis fisicalistas–, Feigl aventuraba que "[…] aquello de lo que se tiene experiencia y (en el caso de los seres humanos) es conocido experiencialmente […] es idéntico al objeto del conocimiento por descripción […] proporcionado en primer lugar por la teoría de la conducta molar, y éste a su vez es idéntico con lo que la ciencia de la neurofisiología describe […] como procesos que ocurren en el sistema nervioso central, quizá especialmente en el córtex cerebral" (Feigl 1958: 446)[83].

Convencido, en cambio, de que una teoría rigurosamente materialista de la mente no tenía por qué verse forzada a comulgar con esa doble tesis de identidad (experiencia = conducta, conducta = actividad nerviosa) –o triple, si, por transitividad, contamos también con que experiencia = actividad nerviosa–, David M. Armstrong (1968) daría por fin forma madura a la teoría de identidad psicofísica postulando su validez para todo el ámbito de lo psicológico: no sólo las sensaciones, sino también creencias o deseos, serían ni más ni menos que estados fisiológicos. De esa manera, Armstrong dotaba a lo mental tanto de la eficacia causal como de la existencia categórica de las que Ryle había querido privarlo y admitía, enfrentándose cara a cara con las tesis de Ryle, que las disposiciones causan sus propias manifestaciones. Pero como bien ha sabido ver Place (1999: 393), eso no es lo único que contribuye a la pujanza de la propuesta de Armstrong, que además "[…] restores the unity of the mental and thus the possibility of being able to define its essence" –posibilidad que se antoja irrenunciable para, por ejemplo, Block (2007b: 9, *infra*)–, y, de paso, ofrece "[…] justification for the philosopher to become involved in the new scientific disciplines or artificial intelligence and cognitive science" (Place 1999: 393), quebrando los recelos ante la investigación científica que Ryle –y también a veces el

[83] *Cf.* Rodríguez y Chacón 2001: 101-102, así como Pujadas 2002: 62-63, para una valoración crítica del papel del eslabón conductual en Feigl.

propio Place– habían dejado entrever. Pese a esos méritos de la teoría de Armstrong, Place se ha mantenido fiel a la idea de que una disposición y sus estructuras o mecanismos subyacentes son "[…] 'distinct existences,' to use Hume's phrase, in which the structure stands to cause as the disposition to effect" (Place 1999: 393); "[…t]o that extent" –concluye el propio Place– "I am still a Rylean"[84].

Ahora bien, no es difícil darse cuenta de que la tesis de identidad psicofísica, enunciada con alcance de tipos, es extremadamente ambiciosa incluso en sus versiones más modestas, como la de Place (1956). Los mismos motivos que la harían aparecer a los ojos de Popper como un impecable espécimen de hipótesis científica –los vastísimos flancos de su vulnerabilidad–, la convierten en la enseña de un proyecto de investigación desesperanzado. Bastaría con hallar evidencia sólida de la existencia de un solo sujeto que se encontrara en un determinado estado mental y no en el estado físico que la teoría dicte para que la tesis de identidad psicofísica de tipos cayera en lo relativo a ese tipo de estado psicológico y, dada su severa formulación, para que se desmoronara *ipso facto* como concepción general de la naturaleza de lo mental –aunque no se tratara, como vendría a apuntar Putnam (1967a: 228, *infra*), más que de un pulpo hambriento cuyo estado cerebral no fuera equiparable al que se registra en nosotros cuando sentimos hambre. Con sólo que uno de los sujetos cuya actividad cerebral asociada a un determinado tipo de estado mental hemos ido laboriosamente analizando mostrara un patrón que *fisiológicamente* –en última instancia, físicamente– no pudiéramos considerar del mismo tipo que el que hubiéramos registrado en todos los demás, la hipótesis se vendría abajo, y nos veríamos obligados a depurarla y volver a empezar. Pero la tarea de refinado de nuestra hipótesis se nos haría cada vez más exigente, pues lo que tendríamos que aislar es ni más ni menos que unas condiciones físicas que se dieran en todos y cada uno de los sujetos que albergaran el estado mental objeto de nuestro estudio, y que no se dieran en ningún sujeto que no atravesara dicho estado mental. Si al menos pudiéramos concitar algunos ejemplos paradigmáticos de identidades psicofísicas de tipos irreprochablemente consolidadas –ejemplos de esos de los que, por hacer buena la descripción de Kuhn (1962, 1969) de una matriz disciplinar, pudiéramos cargar los libros de texto para adoctrinamiento de los estudiantes–, quizá el horizonte se nos hiciera menos desolador, pero no es descabellado aventurar que a quien espigara el vergel de las ciencias del cerebro buscando eso y no otra cosa entre sus frutos le parecería hallarse en mitad de un páramo.

Además, adherirnos a una formulación de la tesis de identidad psicofísica con alcance de tipos nos forzaba, al menos en principio, a guarnecer la identificación propuesta para cualquier variedad de sujeto con que pudiéramos topar: personas, animales de cualquier otra especie, máquinas, acaso algún ser vivo desarrollado sobre una bioquímica diferente de la que conocemos, o incluso seres de naturaleza espiritual si los hubiera. Tan exuberante proliferación de refutaciones tentativas se

[84] Aunque no está del todo claro, como se ha visto *supra*, que la posición de Ryle no se acercara más a la de Price (1953), que Geach (1957) impugnaría.

hacía más estrepitosa cuanto más heterogéneo el inventario de sujetos imaginados, pero seguramente –como se verá: *cf.* Bickle (2006: §1.5, *infra*)– fuera más grave en el caso si se quiere más vulgar, el de varios sujetos de la misma especie, o incluso el de uno mismo a lo largo del tiempo. De cualquier manera, la abrumadora distancia entre lo que parecía exigible a la hipótesis de identidad psicofísica de tipos y lo que ésta podía aducir en su propia defensa acabaría llevando a que se perfilara como más razonable una cuidadosa debilitación de la hipótesis.

Aseverar que cada estado mental concreto –cada instancia, cada caso– no es más que un estado cerebral concreto resulta, con creces, mucho menos comprometido, pero no menos materialista, que flanquear el signo de identidad con tipos de estados y no con casos. De hecho, el rigor materialista de la propuesta parece intachable: una vez concedido que cuando *a* contempla –digamos– las manchas de azul Patinir que vertebran sobre el lienzo las estribaciones de una cordillera lejana, su estado mental es un determinado estado físico de su sistema nervioso, y que el estado mental de *b* al contemplar la misma escena también es un determinado estado físico de su sistema nervioso, resulta prejuicioso afirmar que es más estrictamente materialista quien asegura *además* que el estado físico de *a* y el estado físico de *b* pertenecen a un mismo tipo de estados que hemos fijado atendiendo únicamente a propiedades físicas, que incluye a todos aquellos estados que consisten en la experiencia ocasionada por la contemplación de unas pinceladas de azul Patinir, y que excluye a todos los que no pudiéramos describir así. Al hilo de los planteamientos de Fodor (1974): el fisicalismo de tipos no es una variedad más escrupulosamente materialista del fisicalismo de instancias, sino la conjunción de fisicalismo de instancias con la improbable tesis epistemológica de que una física idealmente madura contaría con un predicado definitorio de una clase natural coextensivo con cada uno de los predicados definitorios de clases naturales que pudieran formularse en cualquier ciencia especial idealmente madura; "reduccionismo" sería, así pues, un nombre más perspicuo que "fisicalismo de tipos"[85].

[85] La puntualización de Fodor hace comprensible que una distensión de los mecanismos de reducción interteórica pueda recoger la misma mies que la debilitación del alcance de la tesis de identidad psicofísica, aunque sea –cabe sospechar– no faenando a plena luz, sino con un grado de transparencia algo menor. No es raro, así pues, que los intentos de recuperar la tesis de identidad con alcance de tipos, motivados por la discrepancia con la interpretación antirreduccionista del funcionalismo, hayan ido a menudo acompañados de una distensión de los cánones que definen el proceso de reducción teórica, como sucede por ejemplo en Hooker (1981), Churchland (1989a) o Bickle (1998). De lo que se trata, en estos casos, es de poner en pie una noción de reducción que no sea tan exigente como para dar al traste con toda posibilidad de reducir la teoría psicológica a teoría neurológica, permitiendo cierta laxitud en el grado en que los enunciados de la teoría reducida deben poder ser expresables mediante el vocabulario de la teoría reductora. Como deja dicho Liz (1995: 224-225), "[…l]a reducción de teorías se convierte, con todo esto, en cierto tipo de *explicación* de la teoría reducida en términos de la teoría reductora más que en un asunto de simple derivabilidad lógica". No es raro que esto sea así porque, al fin y al cabo, debilitar la noción de reducción teórica que se halla encardinada en la tesis de identidad psicofísica es una manera de mitigar las exigencias que penden sobre ésta sin renunciar a la radicalidad que le otorga el alcance de tipos. Pero sobreviene entonces el recelo de que en los

Formalmente, la operación que franquea el tránsito hacia una tesis de identidad con alcance de casos consiste sólo en anteponer el cuantificador universal referido a los posibles sujetos de que se predica la propiedad mental al cuantificador existencial referido a la propiedad física con la que se identifica ésta, de manera que ya no leemos "para toda propiedad mental *M* existe una propiedad física *F* tal que para todo *x*…", sino "para toda propiedad mental *M* y para todo *x* existe una propiedad física *F* tal que…": *(M)(x)(∃F)(MxFx)* (Liz 1995: 223). La intuición, en definitiva, que da fuerza al argumento es la siguiente: no nos extrañaría descubrir que dos sujetos (o un mismo sujeto en distintos momentos) se encuentran en sendos estados físicos que clasificados de acuerdo con criterios de orden psicológico resultan ser del mismo tipo, pero que no hay manera de encajar como pertenecientes al mismo tipo de estado si los criterios que se emplean son estrictamente físicos. En realidad, si la intuición es certera, casi nos sorprendería, a la luz de lo que sabemos acerca de la estructura y función del sistema nervioso, encontrarnos que tal cosa resultara imposible. Según lo expresan Chacón y Rodríguez (2001: 107), "[…e]s fácilmente concebible y muy probable que una misma actividad mental no requiera corresponderse con idénticas características de un mismo proceso orgánico".

Nadar y guardar la ropa: conductismo, fisicalismo y teoría de autómatas

Quién sabe si sería fácil, en aquel contexto, apreciar la manera en que las corrientes que agitaban el cauce de la teoría de identidad psicofísica resonaban en el audaz análisis de Putnam (1967b) de los conceptos centrales de la teoría de autómatas. Lo cierto es que la distinción entre una versión dura de la tesis de identidad psicofísica, en la que a la relación de identidad se le concediera un alcance de tipos, y una versión más dúctil, en la que se le diera sólo un alcance de casos, o instancias, enlazaba suavemente con la necesidad de fraguar una concepción de lo mental que pudiera albergar no sólo a la mente humana, o quizá de otros animales más o menos cercanos a nosotros en términos evolutivos, sino también a esas mentes artificiales que cada vez más nítidamente se adivinaban en el horizonte –o incluso a las mentes extraterrestres que colonizaban, casi siempre amenazantes, vastos territorios de la cultura popular. Como ha recordado Fodor (1985):

> It looked, in the early 1960s, as though anybody who wanted psychology to be compatible with a physicalistic ontology had a choice between some or other kind of *behaviorism* and some or other kind of *property-identity theory*. For a variety of reasons

---

vericuetos de este distendido concepto de explicación de una cierta lectura de una teoría a la luz de otra más básica vuelvan a ocultarse las cláusulas *caeteris paribus* de innegable contenido psicológico que el conductismo lógico parecía pretender que pasáramos por alto. Desde luego, no querríamos haber caminado tanto para regresar a ese mismo paraje.

> neither of these options seemed very satisfactory (in fact, they still don't), so a small tempest brewed in the philosophical teapot.
>
> What came of it was a new account of the type/token relation for psychological states: psychological-state tokens were to be assigned to psychological state-types *solely* by reference to their causal relations to proximal stimuli ('inputs'), to proximal responses ('outputs'), and to one another. The advertising claimed two notable virtues for this theory: first, it was *compatible* with physicalism in that it permitted tokenings of psychological states to be identical to tokenings of physical states (and thus to enjoy whatever causal properties physical states are supposed to have). Second, it permitted tokens of one and the same psychological-state type to differ arbitrarily in their physical kind. This comforted the emerging intuition that the natural domain for psychological theory might be physically heterogeneous, including a motley of people, animals, Martians (always, in the philosophical literature, assumed to be silicon based), and computing machines. (Fodor 1985: 15)

El desolado paisaje en que afloró, entre los estertores del conductismo, y la promesa de autonomía explicativa para la psicología sin detrimento de su compromiso fisicalista, darían cuenta según el relato de Fodor de la buena acogida prodigada al funcionalismo entre quienes se dedicaban entonces a la investigación experimental en psicología:

> Functionalism, so construed, was greeted with audible joy by the new breed of 'Cognitive Scientists' and has clearly become the received ontological doctrine in that discipline. For, if Functionalism is true, then there is plausibly a *level of explanation* between commonsense belief/desire psychology, on the one hand, and neurological (circuit-theoretic; generally 'hard-science') explanations on the other. "Cognitive Scientists" could plausibly formulate their enterprise as the construction of theories pitched at that level. Moreover, it was possible to tell a reasonable and esthetically gratifying story about the relations *between* the levels: commonsense belief/desire explanations *reduce* to explanations articulated in terms of functional states (at least the true ones do) because, according to Functionalism, beliefs and desires *are* functional states. And, for each (true) psychological explanation, there will be a corresponding story, to be told in hard-science terms, about how the functional states that it postulates are "realized" in the system under study. Many different hard-science stories may correspond to one and the same functional explanation since, as we saw, the criteria for the tokening of functional states abstract from the physical character of the tokens. (Fodor 1985: 15)

La reflexión en torno a la naturaleza del vocabulario teórico empleado para describir el funcionamiento de automátas abstractos fue el caladero donde Putnam encontró las claves de esta nueva forma de pensar acerca de lo mental –no sin cierta renuencia, pues el propio Putnam (1960) rechazaba buena parte de las conclusiones que abrazaría en Putnam (1967a, 1967b), y que luego, al menos desde Putnam (1988), volvería a denostar. Comoquiera que fuese:

> […i]t seems that to know for certain that a human being has a particular belief, or preference, or whatever, involves knowing something about the functional organization of the human being. As applied to Turing Machines, the functional organization is given by the machine table. A description of the functional organization of a human being

> might well be something quite different and more complicated. But the important thing is that descriptions of the functional organization of a system are logically different in kind either from descriptions of its physical-chemical composition or from descriptions of its actual and potential behavior. (Putnam 1967b: 424)

Así pues, las operaciones de la mente, como las de una máquina de Turing, quedarían descritas, al menos en una primera aproximación por el programa o "tabla de maquina", un conjunto finito de instrucciones[86] que especifican las eferencias y estados internos resultantes de cada combinación de aferencias y estados internos previos, para un número también finito de estados internos, aferencias y eferencias. En efecto, lo que solemos llamar máquina de Turing —el propio Turing (1948: 7) las llamaba, más modestamente, "logical computing machines"– es una especificación de la estructura abstracta de una máquina descrita por primera vez en Turing (1936, 1937): se trata, si se quiere, de una *máquina abstracta*, en el bien entendido de que tal cosa no es un ingenio construido con tuercas y tornillos, sino un objeto lógico-matemático. Las capacidades de la máquina son escasas pero inequívocamente definidas. Una máquina de Turing está dotada para seguir instrucciones del tipo {$Estado_i$, *Símbolo*, $Estado_f$, *Acción*}, que llamamos reglas de transición. Los dos primeros términos de cada regla especifican las condiciones en las que ésta es de aplicación: cuando la máquina se encuentre en un determinado $Estado_i$ –estado inicial– y detecte un determinado *Símbolo* (digamos, convencionalmente, *0* ó *1*). Los dos últimos términos de cada regla especifican, en cambio, lo que la máquina hará si se dan dichas condiciones: pasará a encontrarse en el $Estado_f$ –final– y ejecutará la *Acción* establecida. Dichas acciones pueden ser sólo cuatro: escribir *0*, escribir *1*, desplazarse hasta el siguiente símbolo, o desplazarse hasta el anterior[87]. Concluida la aplicación de una regla, el ciclo se repite, de modo que el estado final al que ha pasado la máquina es ahora su estado inicial y el símbolo que ha quedado consignado como resultado de la acción especificada –ya sea el que la máquina ha escrito o al que se ha desplazado– es ahora el símbolo que determinará qué acción se aplica. Cuando hay

---

[86] *Sistemas de producción* habrían de denominarse rutinariamente, no sin cierta imprecisión, desde que Newell y Simon (1972) presentaran su modelo global de resolución de problemas con control distribuido.

[87] Resulta evidente, entonces, que las reglas de transición podrían también describirse como quíntuplas {$Estado_i$, $Símbolo_i$, $Estado_f$, $Símbolo_f$, *Desplazamiento*}, que es de hecho el formalismo elegido originalmente por Turing (1936, 1937). No es menos evidente que, comparada con la formulación mediante cuádruplas {$Estado_i$, *Símbolo*, $Estado_f$, *Acción*}, las reglas de cinco términos dotan a la máquina de mayor velocidad de trabajo –pues puede en un único paso registrar un nuevo símbolo y desplazarse al anterior o al siguiente, tarea para la que precisaba dos pasos empleando reglas de cuatro términos–, pero no la revisten de ninguna capacidad básica adicional. Dicho de otro modo, toda tarea que pueda hacer una máquina descrita mediante quíntuplas podrá desempeñarla también una descrita, más parsimoniosamente, mediante cuádruplas.

Sea como sea, la máquina cuenta con tres acciones susceptibles de programación: el cambio de estado, el cambio de símbolo, y el cambio de posición. Además, evidentemente, ha de contar con algunas otras capacidades elementales, tales como responder discriminativamente a sus propios estados, o albergar estados diferenciados.

una regla aplicable, se aplica. Cuando no hay exactamente una regla de transición aplicable –ni más ni menos que una–, la máquina se detiene[88].

Vista desde una perspectiva, si se quiere, algo más escorada hacia lo estructural, una máquina de Turing está formada por una tabla de máquina –el inventario de reglas de transición o, más condensadamente, una función de transición–, un dispositivo de lectura y escritura –que detecta símbolos y, llegado el caso, los reemplaza por otros–, y un dispositivo de memoria, que Turing, en consonancia con la tecnología de su tiempo, imaginó como una cinta continua, dividida en celdas –una celda por símbolo, un símbolo por celda si contamos el espacio vacío como un símbolo–, que la máquina puede desplazar de acuerdo con las instrucciones –o, equivalentemente, desplazarse por ella. La contrapartida de unas capacidades básicas tan sencillas es una doble idealización de, digamos, las condiciones de trabajo de la máquina: no debemos exigirle ningún plazo para completar su tarea, y no debemos imponerle ninguna limitación de memoria. Es decir: debe contar con un tiempo infinito y con una cinta también infinita –si es que es mediante una cinta, según ideó Turing, como se implementa la memoria de la máquina.

Así, otra forma esquemática para cada una de las instrucciones que conforman la tabla de máquina, equivalente a la cuádrupla {$Estado_i$, *Símbolo*, $Estado_f$, *Acción*} pero en la que los ecos psicológicos se hacen ya algo más diáfanos, vendría a ser: "Si el sistema se encuentra en el estado S_i y recibe la aferencia I_j, pasará al estado S_k y emitirá la eferencia O_l". Pues bien, la audaz intuición de Putnam fue que nuestros propios estados mentales –nuestras creencias, deseos, sentimientos…– no serían, entonces, sino los estados S_1 … S_n que vendrían mencionados en nuestra hipotética tabla de máquina[89]; de ahí que la expresión "funcionalismo de tabla de máquina" pronto arraigara como topónimo de la ubérrima región teórica descubierta por Turing, y que Putnam nos invitaba a explorar.

Toda vez que una máquina de Turing es un autómata abstracto, que puede en manos del ingeniero materializarse de maneras tan dispares desde un punto de vista físico como queramos imaginar, en el camino abierto por Putnam se vislumbraba además un horizonte de autonomía para la explicación psicológica, que hemos de recorrer con más calma *infra*. La psicología, desde luego, parecía presta a ejercer una soberanía respecto al estudio del sistema nervioso –la misma que desligaba a la teoría de automátas del estudio de mecanismos concretos– que había anhelado desde su

[88] Esto hace de la máquina de Turing un autómata determinista; si especificáramos diferentes probabilidades para cada transición nos hallaríamos ante un autómata probabilista, que no se detendría necesariamente cuando hubiera más de una regla de transición aplicable.

[89] Es posible también dar a la analogía entre estados mentales y estados de tabla de máquina un referente más lato, haciendo corresponder a cada estado mental de un organismo una *descripción instantánea* de la máquina de Turing con la que se identifica. Una descripción instantánea de una máquina de Turing es el trío formado por el estado en que se encuentra la máquina, el listado de símbolos registrados en la memoria –es decir, el contenido completo de la cinta–, y la posición de registro de la máquina –la celdilla de la cinta sobre la que se encuentra el cabezal de lectura y escritura. La analogía parece perder así, sin embargo, parte de su seducción intuitiva.

propia fundación como disciplina científica. Esos anhelos cristalizarían en Fodor (1974), donde la unidad de la ciencia quedaba perfilada como una veleidad positivista y la independencia de las ciencias especiales, cada una con su vocabulario explicativo intransferible, se erigía como un signo irrevocable de que los tiempos, también en esto, estaban cambiando.

Como acertadamente ha sabido resumir Block (1996), las reflexiones de Putnam guardaban al menos los siguientes réditos:

> (1) According to functionalism, the nature of a mental state is just like the nature of an automaton state –that is, [it is] constituted by its relations to other states and to inputs and outputs. […] (2) […] Mental states can be totally characterized in terms that involve only logicomathematical language and terms for input signals and behavioral outputs. […] (3) […A]lthough functionalism characterizes the mental in nonmental terms, it does so only by quantifying over realizations of mental states […]. (4) One functional state can be realized in different ways. […] (5) […O]ne physical state can realize different functional states in different machines. […] (6) […] For this reason, functionalism shows that physicalism is false: if a creature without a brain can think, thinking cannot be a brain state. (Block 1996: 17-18)

También apunta Block (1996: 17) que, mientras que (2) parece satisfacer el venerable desiderátum conductista de desarrollar una descripción completa de lo mental en términos no mentalistas –e incluso, suponiendo que fijáramos ese criterio como demarcación del conductismo, parecería convertir al funcionalismo en una variedad suya (Block 1980a: 34)–, (3), no obstante, nos muestra que la apariencia es ilusoria. Más terminante se mostraría Fodor (1981a: 10), quien tilda de "[…] more than mildly perverse […]" la lectura del funcionalismo según la cual éste no es sino "[…] a liberated form of behaviorism […]" (*cf*. Leahey 2005: 396, *supra*)[90]. Lo que hace perversa esa interpretación es, de acuerdo con Fodor, que funcionalismo y conductismo únicamente comparten "[…] the relational construal of mental properties", mientras que los distancian "[…] striking differences" en casi cualquier otro ámbito: el funcionalismo no aspira a la reducción ni a la eliminación del discurso psicológico, respalda al pie de la letra la atribución de papeles causales a estados mentales particulares y avala así el concepto de proceso psicológico, etc.

Pero es innegable, *pace* Fodor, que la promesa de una depuración del vocabulario subjetivo, largamente acariciada por los conductistas, está presente en el impulso inicial del funcionalismo tanto como pueda decirse que lo está –por valernos

[90] Asegura Shoemaker, por ejemplo, que es posible interpretar el funcionalismo como la doctrina de que existe un determinado procedimiento capaz de "eliminar" los términos referidos a estados mentales de las explicaciones psicológicas (Shoemaker 1975: 306-307); Block (1978: 64) citaba estas consideraciones de Shoemaker como ejemplo de comunión entre funcionalistas y conductistas. También Dennett (1978b: 62) –después de todo, un discípulo de Ryle– hace notar que la validez del empleo de vocabulario intencional en la explicación de la conducta viene supeditada a la disponibilidad de un método para finalmente "librarse de" dicho vocabulario. Como hace notar Gardner (1985: 97), no es casual que tan tímida concesión surja en el contexto de una valoración de las aportaciones teóricas de Skinner; *cf*. también Block (1978: 78, *infra*).

del modismo acuñado por Skinner (1987: 783, *infra*)– la de su restauración. La propuesta de especificación funcional del dolor bosquejada por Putnam es poco menos que una declaración de principios:

> […W]e *can* specify the functional state with which we propose to identify pain, at least roughly, without using the notion of pain. Namely, the functional state we have in mind is the state of receiving sensory inputs which play a certain role in the Functional Organization of the organism. This role is characterized, at least partially, by the fact that the sense organs responsible for the inputs in question are organs whose function is to detect damage to the body, or dangerous extremes of temperature, pressure, etc., and by the fact that the "input" themselves, whatever their physical realization, represent a condition that the organism assigns a high disvalue to. (Putnam 1967a: 229)

En efecto, el funcionalismo considera los estados de tabla de máquina de un autómata como "estados de segundo orden", definidos por su posesión de propiedades físicas (de primer orden) en determinadas relaciones recíprocas. Estas propiedades de primer orden se consideran implementaciones (instanciaciones, realizaciones) de las propiedades funcionales, del mismo modo que los estados físicos definidos por la posesión de dichas propiedades de primer orden, sin consideración de relaciones funcionales, se consideran implementaciones de los estados funcionales. La teorización funcionalista, así pues, implica cuantificación sobre estados y propiedades físicas consideradas como implementaciones de estados y propiedades psicológicas, aunque no las mencione expresamente en tanto que tales implementaciones.

En el autómata más simple considerado por Block, capaz tan sólo de determinar si el número de *inputs* iguales que ha recibido es par o impar, el primero de sus dos estados de tabla de máquina posibles, S_1, se define –con cierta laxitud en cuanto a la distinción entre estados y propiedades– de la siguiente manera:

> Being in S_1 = Being an x such that $\exists P \exists Q$ [If x is in P and gets a '1' input, then it goes into Q and emits "Odd"; if x is in Q and gets a '1' input, it goes into P and emits "Even" & x is in P]. (Block 1996: 17)

La definición de S_2 sería idéntica, salvo porque establecería al final que x se encuentra en (un estado caracterizado por la posesión de) Q. Ahora bien, P y Q son en la definición variables existencialmente cuantificadas que se refieren a cualesquiera propiedades físicas (o estados caracterizados por su posesión) sirvan como implementación de los estados funcionales del sistema, S_1 y S_2. En la definición funcionalista de un estado mental, por consiguiente, las variables cuantificadas se referirán a estados o propiedades físicas en tanto que implementación de estados o propiedades psicológicas. Para el caso del dolor, por ejemplo, el procedimiento simplificado de formulación de enunciados de identidad teórica seguido por Block, que goza hoy por hoy de cierto carácter canónico, quedaría como sigue:

> Let T be a psychological theory (of either common sense or scientific psychology) that tells us (among other things) the relations among pains, other mental states, sensory inputs, and behavioral outputs. Reformulate T so that it is a single conjunctive sentence with all mental state terms as singular terms […]. Let T so reformulated be written as
>
> $T(S_1 \ldots S_n)$
>
> where $S_1 \ldots S_n$ are terms that designate mental states. Replace each mental state term with a variable and prefix existential quantifiers to form the Ramsey sentence of the theory
>
> $\exists x_1 \ldots x_n\, T(x_1 \ldots x_n)$.
>
> […] Now, if x_i is the variable that replaced "pain", we can define "pain" as follows:
>
> y has pain if and only if $\exists x_1 \ldots x_n\, [T(x_1 \ldots x_n)$ & y has $x_i]$.
>
> […] Then relative to the theory T, pain can be identified with the property expressed by the predicate
>
> $\exists x_1 \ldots x_n\, [T(x_1 \ldots x_n)$ & y has $x_i]$. (Block 1980a: 30-31)

Desde este punto de vista, no es difícil hacer patente que el uso del formalismo de Ramsey (1931) que antes que Block pusiera en práctica Lewis (1972)[91] tenía una viva fuente de inspiración en el esfuerzo de Smart (1959, *infra*) por desarrollar un análisis temáticamente neutral de los informes de sensaciones[92]. Bajo cierta lectura del trabajo de Ramsey, en efecto, cabría interpretar que reformular una teoría como una cuantificación existencial sobre variables, empleando lo que Hempel (1958) bautizó como "enunciados de Ramsey", depura a la teoría de compromisos ontológicos –por usar vocabulario prestado de Kuhn (1962). Así, el enunciado de Ramsey de una teoría se convertiría en el trasunto de los sistemas deductivos puramente formales que Schlick (1918/1925: 37) identificaba con las teorías científicas: entidades abstractas que "[…] flotan libremente" sin que ninguno de los conceptos que las forman "[…] designen nada real", que afirman, como mucho, que *si* hubiera en la realidad algo que satisficiera los axiomas de la teoría, *entonces* la teoría nos daría una descripción correcta de ello. Pero la posición de Ramsey bien puede entenderse como opuesta a la de Schlick en este punto: de hecho, su oposición a la concepción formalista de las matemáticas de Hilbert era notoria (*cf*. por ejemplo Ramsey 1931: 68), y se mostraba aún más reacio a aceptar el formalismo como análisis del conocimiento científico en general. Como acertadamente apunta Psillos (2006: 70),

[91] El procedimiento, de hecho, es conocido también como método Ramsey-Lewis de definición de términos teóricos, o a veces como método Carnap-Ramsey-Lewis: Rudolf Carnap había presentado una propuesta parecida en 1955, antes, por tanto, de que pudiera leer *The Theoretician's Dilemma*, donde Hempel reivindicaba las hasta entonces poco conocidas aportaciones de Ramsey (*cf*. Psillos 1999, 2006). El trabajo de Carnap se basaba en los resultados de Craig (1953) sobre axiomatización recursiva, pero desembocaba en un régimen de introducción de términos teóricos muy similar al perfilado por Ramsey.

[92] Una observación parecida puede verse en Pujadas (2002: 46).

esa extensión del programa de Hilbert a todo el ámbito de la ciencia era precisamente el proyecto de Schlick. Tanto la cercanía como la distancia entre Schlick y Ramsey resultan desde esa perspectiva transparentes: si bien ambos tratan los términos teóricos como variables,

> […] in opposition to Schlick, […] [Ramsey] thinks that advocating an empirical theory carries with it a claim of *realization* (and not just an if-then claim): *there are* entities which satisfy the theory. This is captured by the existential quantifiers with which the theory is prefixed (Psillos 2006: 71).

Desde luego, que estuviera en el ánimo de Ramsey una interpretación siquiera moderadamente realista de la referencia de las teorías científicas, contraria a la posición instrumentalista que a menudo se le ha atribuido (*cf*., por ejemplo, Sahlin 1990), no entraña que tal sea la interpretación correcta; incluso en lo que concierne a los compromisos ontológicos de su propia reconstrucción de la naturaleza de las teorías, el diagnóstico de Ramsey podría ser erróneo. Sin embargo, basta realizar una somera cata en dicha reconstrucción para ver que las conclusiones realistas son, cuando menos, coherentes con la idea de las teorías científicas que Ramsey dejó articulada antes de su prematura muerte. Una teoría –dice Ramsey– tiene por núcleo un conjunto de proposiciones –predicados o funciones que operan sobre constantes individuales– mediante los que se describen hechos empíricos que requieren explicación: se trata del *sistema primario* de la teoría. Sobre ese núcleo se yergue el *sistema secundario*, en el que se introducen proposiciones formadas con nuevos predicados o funciones, de naturaleza teórica, y con las constantes del sistema primario, siempre bajo el dominio de un conjunto finito de axiomas. Ambos sistemas, por último, están ligados por un *diccionario* que define los predicados o funciones del sistema primario en términos de predicados o funciones del sistema secundario (Ramsey 1931: 215). De la conjunción de los axiomas y el diccionario se derivan proposiciones particulares –"consecuencias" de la teoría– y también generales –"leyes" [93].

Una pregunta legítima, entonces, es si el sistema secundario aporta contenido alguno a la teoría que no estuviera ya incorporado en el sistema primario. Dado que Ramsey (1931: 219) argumenta que todo cuanto puede expresarse con los recursos

[93] El procedimiento simplificado que describe Block (1980a: 30-31, *supra*) está, así pues, *crudamente* simplificado: en el enunciado de Ramsey de la teoría incluye Block únicamente la conjunción de sus términos teóricos –*i.e.*, las constantes extralógicas del sistema secundario de la teoría–, mientras que Ramsey incorpora además las constantes extralógicas del conjunto de axiomas de la teoría, que forma parte también de su sistema secundario, y del diccionario que permite verter el vocabulario del sistema secundario al del sistema primario; todo ello es, entonces, reemplazado por variables que se ligan a cuantificadores existenciales. No es fácil ver como la versión de Block de un enunciado de Ramsey de una teoría podría mantener las propiedades lógicas que hacen interesantes a los enunciados de Ramsey: que el enunciado de Ramsey de T es consecuencia lógica de T, que reproduce su estructura deductiva, sus consecuencias observacionales y por tanto su adecuación empírica, su patrón de relaciones interteóricas, etc. (*cf*. Psillos 2006: 72)

del sistema secundario puede expresarse también sólo con los del primario, y que la estructura íntegra del sistema secundario puede reproducirse por medio de definiciones explícitas en el seno del sistema primario siempre que no se fijen límites a la complejidad de esas definiciones (Ramsey 1931: 220), la respuesta que se perfila parece ser que el sistema secundario no aporta contenido a la teoría, o al menos que no es necesario que así sea, con la consiguiente invitación al instrumentalismo o a un reduccionismo quizá al estilo de Hempel (1935). Pero –piensa Ramsey– otra pregunta legítima es si existen, no obstante, usos legítimos de la teoría que no pasen por la formulación de esas exhaustivas definiciones explícitas: la respuesta de Ramsey es que hay motivos metodológicos que hacen desaconsejable depender sólo de las definiciones explícitas de los predicados o funciones del sistema secundario en términos de predicados o funciones del sistema primario, es decir, tomar dichas definiciones como necesarias y el sistema secundario en cambio como enteramente prescindible: fundamentalmente, la necesidad de que el vocabulario de una teoría pueda comenzar a aplicarse a fenómenos nuevos sin que ello implique forzosamente un cambio de significado de dicho vocabulario (*cf*. Ramsey 1931: 230), así como la necesidad de dar cuenta de la posibilidad de que el referente de un término teórico llegue a ser observado empíricamente y el predicado correspondiente deba abandonar el sistema secundario para incorporarse al primario (*cf*. Ramsey 1931: 262)[94]. De modo que la perspectiva de que el sistema secundario aporte al fin y al cabo contenido a la teoría no queda cerrada, y la idea de teoría de Ramsey resulta ser compatible con alguna suerte de realismo. La clave, claro, radica en preguntarse cuál es el contenido adicional que los términos teóricos del sistema secundario aportan a la teoría; la respuesta, como ha argumentado Psillos (2006: 70), bien podría pasar por que dicho contenido sea, precisamente, un juicio existencial categórico, el que expresan los cuantificadores existenciales en el enunciado de Ramsey de la teoría.

---

[94] Como ocurrió con los *neutrinos* que Wolfgang Pauli postulara en 1930 en su célebre carta a los miembros del seminario de física de Tubinga, y que Cowan y Reines lograron observar en 1956 (Reines y Cowan 1956, Cowan, Reines y Harrison 1956); el ejemplo es un débito de Pujadas (2002: 80). Suele citarse a modo de paradigma de ese proceso el descubrimiento por Watson y Crick (1953) de la estructura helicoidal de las moléculas de ácido desoxirribonucleico (ADN) y el papel de los pares de bases de nucléotidos en la codificación del genoma, que nos dotó de la capacidad de observar los mecanismos mendelianos de la herencia. Aunque la distancia que media entre Mendel y el ADN es sin duda mucho mayor que la que separa a Pauli del neutrino, la analogía ha hecho fortuna en el campo de las ciencias cognitivas: tan es así que lo que impulsa el proyecto de naturalización de la intencionalidad ha sido descrito a veces como "[…] the hope of someone playing Watson and Crick to Fodor's, Putnam's, or Pylyshyn's Mendel" (Horst 1996: 265). La descripción del proceso que da el propio Ramsey, por otro lado, alude a la definición cinética de la temperatura pero también, adelantándose a los tiempos, a Mendel:

> Of course, fictitious, or "occult" qualities may cease to be so as science progresses. *E.g*. heat, the fictitious cause of certain phenomena […] is discovered to consist of the motion of small particles. So perhaps with bacteria and Mendelian characters or genes. This means, of course, that in later theory these parametric [secondary system] functions are replaced by functions of the given [primary] system (Ramsey 1931: 262)

Debe advertirse, sin embargo, que el hecho de que sea sostenible una lectura realista de los enunciados de Ramsey no entraña que no pueda dárseles el uso que suscitó el interés de Lewis (1972). Ciertamente, una cosa es que reformular nuestras teorías científicas al modo articulado por Ramsey nos conduzca a que éstas queden privadas de todo compromiso ontológico –que ni siquiera aseveren la existencia de nada, que carezcan por tanto de contenido categórico, fáctico, que tengamos, a fin de cuentas, dificultades para dotarlas de un valor de verdad definido–, y otra bien distinta es que la teoría quede así depurada del compromiso con la existencia de entidades de una determinada naturaleza –por ejemplo, de naturaleza mental–, que es lo que, en la estela como se ha dicho de Smart (1959), parece interesar a Lewis, y lo que pudiera sustentar la tesis de que el funcionalismo no se aparta en este terreno del conductismo[95]. En efecto, si la teoría cuyo enunciado de Ramsey construimos contuviera entre los predicados de su sistema secundario términos que atribuyeran a entidades no observadas propiedades categóricas –que son estados mentales, que son partículas, etc.–, el enunciado de Ramsey nos permitiría deshacernos de tales aseveraciones y preservar únicamente la afirmación de que existe una clase de entidades que satisface los detalles de la teoría. Pero toda vez que la aproximación funcionalista a la naturaleza de lo mental se caracteriza precisamente por el hecho de que la propia teoría define en términos relacionales las entidades teóricas que postula, someter la teoría al régimen de Ramsey no puede adelgazar aún más sus compromisos ontológicos, puesto que decir sólo que entre las entidades postuladas por la teoría se dan las relaciones que describe la teoría –y, por tanto, que existen tales entidades– no es decir menos que lo que dice la teoría. Antes al contrario, lo que la técnica de Ramsey hace patente es que la posibilidad de desligarse de la afirmación de que los estados funcionales que postula sean estados físicos de tal o cual naturaleza está abierta para el funcionalista: entre las consecuencias lógicas de la interpretación de teorías científicas bajo el formalismo de Ramsey no se cuenta que para cada enunciado de Ramsey haya de existir un único conjunto de entidades que satisfaga la teoría. Que sistemas físicos de muy diferente naturaleza –o incluso, si existieran, sistemas no físicos– podrían quedar descritos por la misma teoría psicológica –entendida, claro como teoría funcional– es precisamente la intuición medular del funcionalismo respecto a lo que se daría en llamar la *realizabilidad múltiple de lo mental*. Bajo el prisma de las reflexiones de Ramsey sobre las teorías, esa intuición toma una forma a primera vista muy diferente: "[…] a theory need not be a definite description to be a) truth-valuable, b) ontically committing, and c) useful" (Psillos 2006: 72).

Dicho de otra manera, si tomamos como *definiendum* un predicado teórico de una hipotética teoría psicológica *T* –que el sujeto *y* padezca dolor, en el ejemplo de

[95] *Cf*. Psillos (2006: 70-71). Ya Maxwell (1962: 16) insistía, por lo demás, en que el procedimiento de Ramsey no servía para la eliminación de los términos téoricos, como fehacientemente habían mostrado Hempel (1958) y Nagel (1950): "[…] if a given theory […] entails that there exist certain kinds of unobservable entities, then the appropriate Ramsey sentence will also entail that there exist the same number of kinds of unobservable entities" (Maxwell 1962: 17).

Block (1980a: 30-31, *supra*)– referido según T a un estado funcional, el *definiens* formulado mediante el procedimiento de Ramsey-Lewis –a saber, $\exists x_1 \ldots x_n$ [$T(x_1 \ldots x_n)$ & y se halla en x_i]– forzosamente contará entre sus términos con variables cuantificadas existencialmente –entre ellas, x_i– cuyas relaciones funcionales según vienen fijadas por la teoría quedan preservadas en el enunciado de Ramsey. No es enteramente cierto, entonces, que la descripción funcionalista de lo mental esté limpia de residuos mentalistas –casi sería más cercano a la verdad afirmar que está limpia de residuos fisicalistas. Lo mental, entonces, no puede depurarse del funcionalismo, en la medida en que éste recurre a la cuantificación sobre variables que reemplazan a términos que designan estados mentales, que de dichas variables se predica en el seno de la teoría un cierto patrón de relaciones funcionales, y que la tesis funcionalista es precisamente que esos estados funcionales no son otra cosa que los estados mentales. En este sentido, el funcionalismo bien puede interpretarse –a pesar de lo que pudiera parecer en una primera lectura del procedimiento de definición de términos teóricos inspirado en Ramsey– como ajeno al prurito conductista contra la postulación de fenómenos mentales; desde luego, la presencia de entidades o fenómenos no observables entre los *designata* de la teoría no es algo que el formalismo de Ramsey nos faculte para evitar. O, como con tono mucho más taxativo concluyera Block (1978: 78), "[…i]t is a simple fallacy to suppose that if each mental term is defined in terms of the others (plus inputs and outputs), then each mental state is defined nonmentalistically".

Las convergencias y divergencias entre el conductismo y el funcionalismo, que resurgen, tal como las venimos bosquejando, al hilo de la caracterización del autómata de determinación de cardinalidad par o impar de una cadena de estímulos, o al hilo del análisis del procedimiento de definición de términos teóricos inspirado en Ramsey, quedaron trazadas con nitidez por Block (1980a). Para el funcionalismo, el diccionario del lenguaje mentalista nunca quedaría completo si, tal como pretendía el conductismo lógico, nos vetáramos incluir –*nota bene*: del lado de las acepciones también, no sólo del de las voces– referencias a estados psicológicos. Con toda la contundencia que hubiera podido tener cabida una vez concedido que se trata al fin y al cabo de una cuestión empírica –pese a la apariencia contraria de los argumentos de Geach (1957) y Chisholm (1957)–, esto es lo que señalaban Block y Fodor (1972a), ciñéndose en su planteamiento al minucioso y demoledor examen de la plausabilidad del conductismo desarrollado por Fodor (1968):

> The fundamental argument against behaviorism is simply that what an organism does or is disposed to do at a given time is a very complicated function of its beliefs and desires together with its current sensory input and memories. It is thus enormously unlikely that it will prove possible to pair behavioral predicates with psychological predicates in the way that behaviorism requires –namely, that, for each type of psychological state, an organism is in that state if and only if a specified behavioral predicate is true of it. (Block y Fodor 1972a: 45)

Eficacia causal, relevancia explicativa, autonomía

De la mano de la controvertida cuestión de si el vocabulario teórico de una psicología madura habría de incorporar términos referidos a estados o procesos mentales que no pudieran trocarse en jerga disposicional, la diferencia de mayor calado entre conductismo y cognitivismo tiene que ver, como venimos viendo, con el estatus causal de lo mental –así se subraya también, naturalmente, en el análisis de Block (1980a).

Ya Lewis (1966: 166) se percató de que una de las ganancias explicativas que podía ofrecer el funcionalismo era la de permitir "[…] que las experiencias sean algo real, y que lo sean también los efectos de sus ocurrencias, así como las causas de sus manifestaciones". No en vano, como ya sabemos, cada una de esas ocurrencias o manifestaciones, así tomadas de una en una, bien puede, de acuerdo con el funcionalismo, no ser más que un complejo estado físico –y Armstrong (1986: 83) había dejado escrito que, a la luz del materialismo que él propugnaba, "[…] causality in the mental sphere is no different from causality in the physical sphere". En cambio, eran sabidas las dificultades del conductismo lógico para proveerse de un análisis disposicional del aspecto puramente experiencial, cualitativo, de las sensaciones –dificultades que Place (1956) había intentado solventar, dando así paso a la teoría de identidad psicofísica. Un razonamiento similar anima a Fodor (1981a) a repartir con espíritu salomónico los aciertos del funcionalismo entre el legado fisicalista y el conductista:

> What central state physicalists seemed to have got right –contra behaviorists– was the ontological autonomy of mental particulars and, of a piece with this, the causal character of mind-body interactions. Whereas, what the behaviorists seemed to have got right –contra the identity theory– was the relational character of mental properties. Functionalism, grounded in the machine analogy, seemed to be able to get both right at once. (Fodor 1981a: 9)

La piedra angular del funcionalismo sería entonces la distinción entre enunciados de identidad con alcance de casos y con alcance de tipos, que, asentada sobre esa "analogía de la máquina", permitiría dar la razón al fisicalista en cuanto a la "autonomía ontológica" de los casos particulares de estados mentales, y al conductista en cuanto al "carácter relacional" de las propiedades según las cuales esos casos particulares se clasifican en tipos[96].

[96] De este modo, como con indisimulada ironía glosa la cuestión Pujadas (2002: 12), el funcionalismo se perfilaría inmodestamente como *Aufhebung* hegeliana de la filosofía de la mente, ya que "[…] superaría los defectos de sus precursores conservando en cambio [sus] virtudes". Ahora en serio: aunque le aguardara una genuina refutación, o tal vez el lento enmohecimiento del olvido, no parece demasiado osado sostener que el giro que el funcionalismo ha imprimido a nuestros modos de concebir lo mental es de una envergadura que no siempre resulta fácil aquilatar. Tan es así que

De las observaciones de Wittgenstein (1953) respecto a que el comportamiento no es efecto de procesos mentales, sino parte del concepto de tales procesos[97], el funcionalismo podría rescatar así la tesis positiva –las conductas que éste origine son parte del concepto de un estado mental, como las eferencias que provoque lo son del concepto de un estado de tabla de máquina–, sin verse por ello forzado a aceptar también la tesis negativa –que los estados mentales no son causas de la conducta. En efecto, la conclusión de que los estados mentales no causan conductas no se sigue de la premisa de que los conceptos con los que clasificamos los estados mentales son de naturaleza relacional e incluyen referencias a las conductas que forman el séquito de sus efectos. Se seguiría, en todo caso, de un razonamiento ilegítimo, a saber: si los estados mentales son causas de la conducta, entonces o son estados del cerebro o son estados espirituales que atañen a un alma cartesiana; nuestros conceptos de estados no son conceptos de estados del cerebro (pues no incluyen referencia alguna a procesos neurofisiológicos y sí, en cambio, a las conductas que suelen acompañarlos); luego los estados mentales, tal como los concebimos, no pueden ser estados del cerebro; tampoco son estados espirituales de un alma cartesiana, luego, *tollendo tollens*, los estados mentales no son causas de la conducta. La primera premisa, que establece que sólo estados del cerebro o estados de un alma incorpórea podrían ejercer como causas de la conducta, es heredera de lo que Fodor (1975: 4) bautizó como "el pecado original de la tradición wittgensteiniana", la confusión entre mentalismo –entendido como anticonductismo– y dualismo (*cf.* Hermoso 2001a: 247). Antes bien, los estados mentales, de acuerdo con la concepción funcionalista, son causados por estímulos y por otros estados mentales, y causan conductas y otros estados mentales[98], *pero además* se identifican mediante conceptos que, en efecto, mencionan precisamente sus causas y sus efectos: es decir, conceptos funcionales, o "unidades de equivalencia funcional" que engloban patrones diversos de actividad

---

Broncano (1995), que no vacila en incluir los desarrollos registrados en filosofía de la mente entre los más importantes del siglo, se reserva una tasación aun más generosa para el funcionalismo:

> […L]as diversas formas de funcionalismo son […] la principal innovación metafísica desde las polémicas del XVIII y XIX, más allá de la encrucijada entre el materialismo y el idealismo. (Broncano 1995: 12)

Mucho más comedido se muestra Martínez-Freire (2001: 93), que se limita a tomar nota de que, dada su reconocida neutralidad ontológica, "[…] el funcionalismo, frente a lo que pudiera pensarse, no constituye una solución al problema mente-cerebro". Claro que es precisamente esa condición de ser algo distinto de una solución al problema mente-cerebro lo que seguramente convierte al funcionalismo en una respuesta genuinamente nueva ante dicho problema. Así se hace patente en la tajante afirmación con la que Putnam (1960: 175) abría los párrafos finales de "Minds and Machines": "[…] it is no longer possible to believe that the mind-body problem is a genuine theoretical problem".

[97] Pero *cf.* Chihara y Fodor (1965: 139).

[98] Lo ha expresado con particular elegancia Jacob (2002: 648):

> If a mental property is "functionalizable", then its instantiation can both be the effect of either some sensory input or the instance of some other mental property and the cause of either the instance of some other mental property or of some behavioral output.

neurológica, o de otra índole (Fodor 1968: *passim*). Así pues, las conclusiones alcanzadas por Fodor al cabo de su detenido estudio de los procedimientos explicativos habituales en la psicología cotidiana y en la psicología científica (Fodor 1965, 1968a, 1968b) no distaban mucho de las de Putnam, aunque su punto de partida no fuera tanto la reflexión sobre los trabajos de Turing sino la controversia con Ryle:

> Functionalism just *is* the doctrine that the psychologist's theoretical taxonomy doesn't need to look "natural" from the point of view of any lower-level science. (Fodor 1985: 16)

El propio Ryle (1949), sin embargo, había insistido incansablemente en que afirmar que un estado de ánimo, por ejemplo, consiste en la disposición a desplegar ciertas conductas en ciertas circunstancias no equivale a afirmar que el estado de ánimo *cause* la emisión de las conductas –eso sería incurrir en la "metáfora paramecánica"–, sino más bien a afirmar que el estado de ánimo *es* la mayor probabilidad de que las conductas aparezcan. Desde ese ángulo contempla la cuestión García-Carpintero (1995) cuando escribe que:

> La cuestión de la eficacia causal de la mente tampoco presenta, aparentemente, dificultades para el conductismo. La eficacia causal de los estados mentales, entendidos como disposiciones, viola la completud del mundo físico en la misma medida en que lo hace suponer a la solubilidad de un terrón de azúcar eficacia causal. (García-Carpintero 1995: 51-52)

Así es, así visto. Pero sólo, claro, si el alcance de la afirmación se acota al caso del conductismo lógico, ryleano –hemos tomado nota de los titubeos con que Watson y Skinner abordan la cuestión del estatus causal de los estados mentales–; e incluso en el seno de la concepción de la mente arbolada por Ryle, sólo en la medida en que la noción de eficacia causal de la que se sirve el argumento es de un género tal que cabe atribuírsela a un estado mental incluso cuando toda relación de causa y efecto que éste pudiera trabar viniera determinada por su naturaleza fisiológica[99]. Si, por el contrario, aspirásemos a una idea de la eficacia causal de lo mental que nos permitiera aseverar que algunos estados mentales entablan relaciones de causa y efecto en virtud de sus propiedades estrictamente psicológicas –relaciones, por decirlo de otra forma, que no entablarían sólo en virtud de sus propiedades físicas– entonces es claro que el conductismo no satisfará esta aspiración. El mismo ejemplo esgrimido por García-Carpintero lo muestra: desde luego, atribuir poderes causales a la solubilidad de un terrón de azúcar violaría flagrantemente el principio de completud del mundo físico –o principio de clausura– en la exacta medida en que esos poderes causales se supusieran independientes de los de la estructura molecular del terrón y sus propiedades. El matiz, por supuesto, no es ajeno al entendimiento de

[99] Es decir, en la medida en que se trata de una noción de eficacia causal que se adhiere al principio de herencia proclamado por Kim (1993a: 355, *infra*) respecto de la relación de realización, como la que daría en defender Liz (1995).

García-Carpintero, que apenas unas páginas después nos recuerda cómo el propio Ryle –nos lo hace ver asimismo Place (1999: 379, *supra*)– interpretaba los condicionales subjuntivos de los enunciados hipotéticos con los que expresaba las disposiciones que suponía idénticas a tal o cual estado mental como meras "autorizaciones para la inferencia", es decir, "[…] en términos sólo epistemológicos, carentes de compromiso ontológico sustancial" (García-Carpintero 1995: 67).

Esta misma diferencia entre relevancia explicativa y eficacia causal es trazada con nitidez por Toribio (1995) en el marco de un intento de evitar que las propiedades semánticas de los estados mentales queden despojadas, en el propio seno del cognitivismo, de todo papel en el dinamismo de la vida mental –como se verá con detenimiento *infra*. Ya Jackson y Pettit (1990) habían distinguido entre eficacia causal y relevancia causal en términos muy parecidos –Toribio (1995: 6) misma lo hace notar–; su conclusión, que Toribio rechaza, es que si bien cabe atribuir relevancia explicativa –causal, en su terminología– a las propiedades psicológicas, atribuirles genuina eficacia contravendría el *principio* de que ésta reside únicamente en las propiedades físicas con las que aquellas guardan una relación de superveniencia.

La posición de Jackson queda particularmente clara en la discusión de la cuestión emprendida por Braddon-Mitchell y Jackson (1996), que plantean como un desafío a la concepción funcionalista de lo mental que ellos mismos auspician, estrechamente ligada al fisicalismo. El papel correspondiente al fundamento categórico de las disposiciones lo desempeñan, tal como estamos comprobando, determinados estados o procesos neurofisiológicos, y lo que está en duda es la relevancia de las propiedades psicológicas –más precisamente: intencionales, semánticas– en el tejido de causas y efectos en el que dicha actividad nerviosa se imbrica. De esta manera, el problema puede hacerse descansar sobre la intuición, tan medular al funcionalismo como a nuestra idea espontánea de lo mental, de que las propiedades psicológicas son causalmente eficaces. Es decir, la intuición según la cual "[…m]y belief's being that there is a tiger nearby, its having that content, is in part what makes my legs move", ante la que se opone de inmediato la incredulidad respecto a la verdadera eficacia de esas propiedades semánticas:

> But where in the complex sequence of neurophysiological states that runs from the sensory input triggered by the light rays from the tiger to the contracting of the muscles that move my body away from the tiger does the content properly figure? Each and every transition from one neurological state to the next will be governed by the relatively intrinsic, neurological features of the various states, not by the highly relational features that figure in the most plausible approaches to content. As it is sometimes put, the mind is a syntactic (neurophysiological) engine, so how can semantic (content) properties be doing any driving? (Braddon-Mitchell y Jackson 1996: 257)

Aunque Braddon-Mitchell y Jackson presentan ésta como la última de tres objeciones que ha de enfrentar el funcionalismo en relación con el estatus de la explicación psicológica, es innegable que las otras dos objeciones son en realidad facetas de la

misma cuestión. En efecto, preguntarnos por la eficacia causal de unas propiedades mentales que se entienden idénticas a determinadas propiedades nerviosas –sobre las cuales parece recaer, pues, todo el quehacer causal– no constituye una tarea conceptual significativamente distinta de la de preguntarnos como la explicación psicológica, articulada en términos de esas propiedades mentales, se encardina con la explicación neurofisiológica, articulada según las propiedades nerviosas y, exhaustiva. Así pues:

> The puzzle is not how neurophysiological states and psychological states can both cause behavior. As "they" are one and the same, nothing is easier. The puzzle is to understand the relationship or connection between the two ways of describing one and the same set of states in such a manner that both ways of describing the same set of states can be seen as causal-explanatory, especially in light of the fact that only one way is in principle complete. (Braddon-Mitchell y Jackson 1996: 257)

De la misma manera, en la medida en que las propiedades psicológicas de las que se trata vienen definidas en términos funcionales –es decir, por referencia a sus causas y efectos, externos o internos–, la cuestión puede quedar replanteada como un problema no ya de compaginación de varios vocabularios explicativos, sino de trivialidad del propio vocabulario psicológico, cuyos conceptos presuntamente explicativos incluirían ya el efecto que aspiran a explicar, arruinando dicha aspiración:

> […A]ccording to […] functionalism […], to ascribe a psychological property to someone is to ascribe a highly relational, causal-functional nature to an internal state in them, and one, moreover, that is in part defined by what it causes. But if it is defined by what it causes, how can it explain in any significant sense what it causes? (Braddon-Mitchell y Jackson 1996: 256)

Bajo esta formulación, la analogía con el asunto del carácter explicativo de las disposiciones resulta tan transparente que, de hecho, Braddon-Mitchell y Jackson toman como punto de partida para abordar esta triple objeción al funcionalismo precisamente un análisis de la explicación disposicional. Dos son, a su juicio, las observaciones cruciales: en primer lugar, que apelar a una disposición como explicación de un determinado suceso no es tan ocioso como pudiera parecer, y, en segundo lugar, que los roles funcionales merced a los que se definen a su entender nuestros estados mentales difieren de disposiciones –digamos– mecánicas como la fragilidad o la solubilidad en razón de su complejidad. Si bien de ello no se rinde cuentas, es de suponer que alguna trabazón de ambas consideraciones debería disipar el temor de que la explicación psicológica resulte inane. No obstante, ha quedado ya anotado como la defensa que esgrime Place (1999, *supra*) de la idea de que la apelación a disposiciones goza de un valor explicativo relevante –que es, en lo esencial, una elaboración de la propuesta, algo menos explícita, de Braddon-Mitchell y Jackson (1996: 257-258)– se vuelve sumamente endeble una vez que reparamos en que subsumir un hecho particular –un cristal que se ha hecho añicos– bajo una

generalización según la cual un desenlace semejante es el habitual en circunstancias semejantes –el cristal es frágil: se rompe en circunstancias semejantes a aquellas en que de hecho se ha roto– no colma nuestras expectativa de comprensión ni en contextos cotidianos ni en contextos científicos. Dicho de otra forma: si sabemos que un cristal se ha roto ya sabemos que es frágil, pero todavía queremos saber por qué se ha roto. Que lo que pretenda explicar la disposición sea un hecho particular –y no una propiedad general, como en el famoso ejemplo del medicucho de Molière y los efectos del opio que, como es preceptivo, también mencionan Braddon-Mitchell y Jackson (1996: 257)– alivia en alguna medida, es innegable, la circularidad de la explicación, pues no se enuncia lo mismo con otras palabras, sino que se incluye un caso particular bajo un principio general. Pero mientras no contemos con una explicación eficiente de los mecanismos que respaldan el principio general y de cómo operan en el caso particular, la circularidad permanece larvada. Por otra parte, conviene conceder que en disposiciones como la fragilidad o la solubilidad podamos definir "[…] a more or less homogeneous class of inputs and outputs", mientras que:

> By contrast, the functional roles definitive of distinct psychological states can share particular inputs and outputs. There is substantial overlap in the definitive inputs and outputs for distinct psychological states. (Braddon-Mitchell y Jackson 1996: 258)

Esto contribuirá sin duda a que identificar la base categórica de dichos roles funcionales –es decir, si el funcionalismo está en lo cierto, de nuestros propios estados mentales– sea enormemente más complejo que hacer lo propio para la base categórica de la fragilidad del cristal. Pero no se puede decir, desde luego, que esto sea un descubrimiento notable, ni tampoco que haga menos trivial el intento de explicar un hecho psicológico mencionando que pertenece al tipo de hechos psicológicos que suelen suceder a aquellos como el que lo ha precedido. Es más: las perspectivas de trivialidad son particularmente amenazantes para Braddon-Mitchell y Jackson, en la medida en que su posición es que los roles funcionales que definen nuestros estados mentales están consignados ya en la psicología de sentido común con la que nos manejamos en la vida cotidiana.

Ante la pregunta por las relaciones entre la explicación psicológica y la explicación neurofisiológica, el posicionamiento de Braddon-Mitchell y Jackson se va haciendo más transparente, y con él también la razón por la que su respuesta a la imputación de trivialidad no termina de resultar satisfactoria. La relación entre la explicación psicológica y la explicación neurofisiológica es, a los ojos de Braddon-Mitchell y Jackson, la de lo incompleto a lo completo, lo incierto a lo cierto, lo confuso a lo claro, o, si se quiere, la parte al todo –siendo así que la parte en cuestión es una parte incompleta, incierta y confusamente definida. El parangón que se fija es el de una explicación mecánica en la que conocemos exactamente la magnitud de las fuerzas que se ejercen, en la misma dirección y sentido contrario, sobre una partícula –éste sería el análogo de la explicación neurofisiológica–, y la misma explicación mecánica del movimiento resultante, pero siéndonos conocido únicamente cuál de

las dos fuerzas es de mayor magnitud. El carácter completo, cierto y preciso de la explicación neurofisiológica, y la carencia de esas cualidades en la explicación psicológica, se decreta sin argumento. La principal ventaja de la explicación psicológica sobre la explicación neurofisiológica es, así las cosas, que "[…] está disponible en la actualidad" (Braddon-Mitchell y Jackson 1996: 262): más que un saber, diríamos, es una ignorancia provisional[100]. Aparte de la ventaja transitoria que le da la relativa pobreza de nuestros conocimientos neurofisiológicos, es poco más lo que puede aportar la explicación psicológica: como mucho la posibilidad de expresar –"de manera no trivial", añaden Braddon-Mitchell y Jackson (1996: 263)– las diversas perturbaciones que probablemente no evitarían que se produjera idéntico fenómeno al que de hecho se ha registrado. Parece obvio que esas diversas perturbaciones podrían quedar también descritas en el vocabulario de una neurofisiología madura, al menos de una como la que imaginan Braddon-Mitchell y Jackson, pero ello no es óbice, a su juicio, para la conclusión de que:

> In this sense, the laws of psychology might be described as autonomous, though we suspect that some advocates of the autonomy of psychology have had something stronger (and more dubious) in mind. (Braddon-Mitchell y Jackson 1996: 263-264)

Es cierto que la idea de una psicología autónoma –es decir, con leyes soberanas, irreductibles a las de otras disciplinas científicas– es mucho más ambiciosa, y es cierto también, por supuesto, que su plausibilidad es dudosa. Pero no es menos cierto que resulta dudoso que la relación entre la psicología y la neurofisiología bosquejada por Braddon-Mitchell y Jackson pueda describirse como "autonomía" *en algún sentido* que no pase por una expurgación casi total de las implicaciones del concepto de autonomía. Antes al contrario, "subordinación" se perfila como un nombre bastante más juicioso para la relación descrita.

Si en el ámbito epistemológico la trivialidad de la explicación psicológica y su subordinación a la neurofisiológica se vuelven ingobernables para Braddon-Mitchell y Jackson, difícilmente cabría esperar que su ensayo de solución al problema que planteábamos en primer lugar –la eficacia causal de lo mental–, perteneciendo éste al ámbito de lo que tradicionalmente consideramos cuestiones ontológicas, resultara convincente. Conviene, a fin de adentrarnos en dicho ensayo, retornar al tratamiento de la explicación disposicional que acaba de quedar expuesto para anotar que:

> […] there is a distinction between a property's being causally explanatory and its being causally efficacious; in particular, second-order properties of having a property that plays a certain causal role are typically causally explanatory, but are not themselves causal. The causing is done by the first-order property that plays the role. (Braddon-Mitchell y Jackson 1996: 264)

[100] La expresión, que ha hecho fortuna, parece provenir del ensayo "Sobre la lógica de las ciencias sociales", de Theodor W. Adorno (1969), incluido en el célebre debate editado bajo el título de *La disputa del positivismo en la sociología alemana*, donde se refiere a la sociología.

Dicho sea de paso, es esto precisamente lo que cabría reprochar a Braddon-Mitchell y Jackson: que su empeño en denominar explicación causal a una explicación –la explicación en términos disposicionales, en términos de propiedades de segundo orden– en la que, según sus propios criterios, ni intervienen propiedades causales –causales son sólo las de primer orden, la base categórica de la disposición– ni se rinde cuenta del mecanismo eficiente –de cómo "the causing is done"– sólo parece fundarse en su empeño por denominar explicación autónoma a una explicación que guarda con aquella respecto a la cual la consideramos autónoma una relación con todas las notas de la de subordinación. De todas formas, la balanza vuelve enseguida a vencerse del lado de la reducción: si las propiedades psicológicas pueden ejercer determinada eficacia causal es, según Braddon-Mitchell y Jackson, porque, después de todo, *no son propiedades funcionales*, sino que son "[…] the kinds of states that satisfy the conditions functionalism tells us determine which psychological state a subject is in" –o, lo que a sus ojos viene a ser lo mismo, porque "[…] type-type identity theory is compatible with functionalism –indeed, is the obvious position for functionalists to hold" (Braddon-Mitchell y Jackson 1996: 265).

Con todo, no cualquier punto de vista que, como el de Ryle o el de Braddon-Mitchell y Jackson, trate de forma más o menos abierta de ceñirse a lo epistemológico –apelando a licencias de inferencia, a la noción de carácter causalmente explicativo en ausencia de eficacia causal, etc.– nos garantiza una idea de eficacia causal afín a la concepción conductista de lo mental, ni a los planteamientos reduccionistas de Braddon-Mitchell y Jackson. La cosa cambia como de la noche al día si no traducimos eficacia causal por autorización para la inferencia, o por relevancia explicativa, o por ventajas explicativas de orden pragmático, sino por irreductibilidad: la relevancia en la explicación causal que el conductismo o la tesis de identidad psicofísica toleran en lo mental está inherentemente supeditada a que podamos, al menos en principio, reemplazar la explicación psicológica así construida por una que no apele sino a conceptos físicos, y la ventaja explicativa que pueda proporcionarnos no puede ser más que, en el mejor de los casos, de orden práctico. En resumen: si bien es cierto que desde un determinado punto de vista epistemológico la eficacia causal de la mente es fácilmente integrable en la ortodoxia del conductismo lógico o de una interpretación reduccionista del funcionalismo, no lo es menos que tan pronto como se trata de forzar algún compromiso ontológico en torno a la noción de eficacia causal, o bien de trasladar la perspectiva epistemológica hacia la idea de irreductibilidad, ni el conductismo ni la lectura fisicalista del funcionalismo puede acompañarnos. La eficacia causal de los estados mentales parece, en suma, esconderse en el trecho que para la explicación psicológica separa una mera relevancia explicativa de orden pragmático de una genuina irreductibilidad epistemológica. De hecho, no son pocos quienes consideran que en esa acaso impractible travesía el funcionalismo quedaría tan desnortado como el conductismo; el propio García-Carpintero insinúa un razonamiento de ese tenor cuando apunta que "[…] ciertas dudas sobre la eficacia causal de las disposiciones" serían parte del núcleo problemático "compartido por el conductismo con su heredero natural, el funcionalismo" (García-Carpintero 1995: 52) –como acaso también lo sería, según venimos de examinar, el problema del retorno de lo mental.

Ab Architae columba lignea: madrugada del autómata

No es extraño que atisbar como algo más que una remota visión la posibilidad de constuir máquinas pensantes precipitara cambios radicales en nuestra forma de concebir el pensamiento, y con ello todo lo mental y su vínculo con lo físico. Sabemos, de hecho, que la idea de que incontables artefactos físicamente heterogéneos pueden llevar a cabo una misma tarea, la de que eso exige abordar su comprensión bajo un prisma rotundamente funcional, e incluso la de que tal perspectiva pudiera conllevar una recuperación del vocabulario mentalista que habíamos tratado de desterrar de la psicología, venían bullendo en los escritos de quienes, ya en pleno fragor del conductismo o antes, se esforzaban en construir mecanismos capaces de simular determinadas conductas animales o humanas.

Acaso en los primeros autómatas que formaron la corte de los milagros de la Ilustración, como el célebre *Canard Digérateur* que en 1739 armara Jacques de Vaucanson, el realismo en la apariencia del animal mecánico fuera uno de los empeños del inventor –así sucede también, de hecho, en las otras máquinas que d'Alembert describe bajo la voz *automate* en la Enciclopedia, desde la paloma mecánica cuya construcción Aulio Gelio atribuía a Arquitas de Tarento[101] hasta los tamborileros o flautistas del propio Vaucanson, que se jactaba, hablando de su celebrado pato, de haber imitado no sólo "[…] cada uno de los Huesos, sino también las Apófisis o Abultamientos de cada Hueso" (Vaucanson 1738)–, pero es obvio que ese propósito pronto quedó eclipsado por el de remedar las funciones propias de los organismos con un ánimo –como lo describe Cordeschi (2002: *xiv*)– "declaradamente no mimético"[102], acaso porque el hechizo de la imitación exacta hubiera sucumbido a la industrialización del proceso. No es cosa que mereciera mayor discusión: "[…u]n modelo conductista de aprendizaje por condicionamiento no tiene que salivar como el perro de Pavlov; más bien, debe capturar algunos de los *rasgos esenciales* del

[101] "[…M]uchos filósofos griegos famosos, y Favorino entre ellos […], escribieron de manera taxativa que un artilugio con la forma de una paloma, hecho de madera por Arquitas, basándose en unos principios racionales y mecánicos, había volado […]" (*Noches áticas* X, XII). No puede dejar de acotar d'Alembert: "[…] supposé que ce pigeon volant ne soit point une fable". Aunque el aire fantástico del relato se acrecienta al advertir que la paloma sobrevolaba Tarento, que habían fundado los cretenses guiados –quería el mito– por Yapige, hijo de Dédalo, no parece que el logro atribuido a Arquitas, o algo parecido, quedara muy alejado de la destreza de los artífices griegos.

[102] Bajo la voz dedicada a instrumentos matemáticos, en cambio, se describían *máquinas de calcular*, como la bosquejada por Leibniz en 1685 o la máquina aritmética en la que poco antes, desde 1642, había estado trabajando el joven Pascal, y que en 1647, obrando ya en su poder un privilegio real para su fabricación, le mostraría a Descartes. Se diría, así pues, que los enciclopedistas no llegaron advertir con claridad el papel que artefactos como la *pascalina* acabarían desempeñando en la cristalización del concepto de autómata y su aplicación a la comprensión del entendimiento humano. Un documentadísimo recuento de las raíces históricas de las modernas máquinas de cómputo puede encontrarse en Guijarro y González (2010).

fenómeno de aprendizaje" –escribe Cordeschi (2002: *xiv*) parafraseando a Clark L. Hull. En los primeros párrafos de "An Imitation of Life", el artículo donde presentaba su *Machina speculatrix*, William Grey Walter (1950: 42) apunta incluso que el hecho de que no se de por contenta con calcar la apariencia externa de un organismo es lo que distingue a una aproximación científica a la simulación de los procesos biológicos de una aproximación mágica o ritual.

Ya desde los primeros años del siglo XX, la viva controversia acerca de la naturaleza de los tropismos que enfrentaba a Jacques Loeb –cuyo reduccionismo mitigaría y radicalizaría a la vez, en distintos sentidos, el manifiesto conductista de su antiguo estudiante, Watson– y a Herbert S. Jennings se había ido extendiendo a la cuestión de si lo que nos enseñaban los autómatas fototrópicos que comenzaban a darse a conocer es que "[…] el mismo método […] de la física es válido para la investigación sobre conducta animal" (Loeb 1900: 240), o más bien que existen generalizaciones relativas a la conducta de "sistemas de acción" –sea éste un protozoo, en cuyo estudio se había especializado Jennings, un humano o una máquina– que "[…] no pueden quedar expresadas en el vocabulario de la física y la química" (Jennings 1910: 354). Lo cerca que el entonces profesor de Zoología en Johns Hopkins se hallaba de las reflexiones de Putnam sobre las máquinas de Turing salta a la vista en el modo en que su razonamiento lo conduce de la premisa de que máquinas construidas con diferentes materiales podrían resultar indiscernibles desde el punto de vista de su conducta, a la conclusión de que la física o la química quedarían inermes para explicar dicha conducta:

> From the same mass of substance we can make many different arrangements or machines, acting in entirely different ways, so that we could never predict the reactions of the machines from a knowledge of the chemical and physical properties of the unarranged substance. From a certain mass of material we could make either a clock or a doorbell or a steel trap or a musical instrument, —and we could easily so arrange these that each would respond in its characteristic way when acted upon by an electric current. We could, moreover, make the same machines, showing the same reactions, from a different kind of material, with different properties. We could then never predict the reactions of these by knowing merely the chemical and physical properties of the material of which they are composed. The specific action of each depends on the specific arrangement of its material.
>
> This is exactly what we find in organisms, including the lowest as well as the highest. (Jennings 1910: 361)

Es decir, tal como cifra Cordeschi (2002: 24) las conclusiones de Jennings, que dadas las "condiciones paralelas" (Jennings 1910: 361) que se dan entre organismos y máquinas,

> […] the behavior of inorganic arrangements or machines and that of living organisms alike can be explained by referring to their *functional organization* rather than to their respective physical and physico-chemical properties.

Más o menos al mismo tiempo, la idea de que máquinas físicamente muy dispares bien podían ejecutar idénticas funciones iba ahondando su huella entre los propios ingenieros. Así, cuando uno de ellos, Silas Bent Russell (1913, *infra*) presenta un artefacto ideado para contrastar las explicaciones de inspiración hidráulica que la neurofisiología de su tiempo solía ofrecer a los fenómenos de condicionamiento, tiene cuidado en señalar que el funcionamiento de su máquina podría perfectamente ser reproducido por otras cuyas operaciones se basaran, por ejemplo, en mecanismos eléctricos y electromagnéticos[103]. No mucho después, Thomas Ross (1938: 185), que llevaba ya algo más de un lustro trabajando en el diseño de máquinas capaces de aprender a recorrer laberintos, podía tomar la observación de Bent Russell como un lugar común, y emplearla para respaldar un proyecto de investigación psicológica centrado en la construcción de modelos mecánicos –el *enfoque del robot*, como lo había bautizado Clark L. Hull (*infra*):

> To find the sufficient condition for learning we should try to make a machine that will learn […]. A very persistent mistaken attitude to work of this sort is the idea that the builder of a machine which will learn must think he has built a mechanism physically like that underlying human or animal learning. Nothing could be further from the truth […]; for no truth is more commonplace in mechanics than that, in general, several alternative mechanisms, differing widely in superficial characteristics and forms of energy utilized, can produce the same end result (Ross 1938: 185 *apud* Cordeschi 2002: 111)

También Grey Walter (1951: 61-62) insistiría en la misma idea al presentar su *Machina docilis*, con la que pretendía agregar cierta capacidad elemental de aprendizaje a las conductas exploratorias exhibidas por la *Machina speculatrix* que había descrito un año antes:

> In the *M*[*achina*] *docilis* model the memory takes the form of a damped oscillation, but it could well be any mechanical, chemical, or electrical process in which stored energy is slowly released, as in the escapement of a watch. It is essential only that the enery should be in such a form that it can be readily available for the final operation.

Era pronto, no obstante, para que cuajara la lectura epistemológica de estas observaciones que arraigaría en el cognitivismo –la autonomía de la explicación psicológica, la falsedad del reduccionismo. Aunque un reduccionista cabal como

---

[103] Parece que la fascinación por las propiedades del agua fue uno de los primeros signos de la vocación de ingeniero de Bent Russell –hermano mayor del pintor Charles M. Russell, que alcanzó cierta fama merced a sus paisajes del Lejano Oeste americano y sus retratos de nativos y *cowboys*–:

> As a boy he was always tinkering with gadgets. Reading somewhere that evaporation causes coolness, he rigged up two umbrellas over his bed, and two watering-pots with counter-weights to sprinkle them and a complicated system of gutters to carry off the excess moisture, which as he learned by experiment didn't evaporate. It took him only two hours to get to bed, and once in he couldn't get out. Long afterwards his wife remarked, 'When I heard about the umbrellas I should, right there and then, have refused to marry him'. (Russell 1957: 15-16)

Meyer apreciara el valor retórico que artefactos como el de Bent Russell podían tener en el litigio contra el vitalismo (*cf*. *infra*), e incluso su interés como recurso pedagógico, les exigía, precisamente en nombre de esa utilidad didáctica, que se despegaran lo menos posible del vocabulario neurológico, limitándose –como apunta Cordeschi (2002: 70)– a analogías simples y directas con el sistema nervioso: que la mera posibilidad de desligarse de ese vocabulario pudiera apremiarnos a una reconsideración de la idea de reducción del vocabulario psicológico al neurológico ni siquiera llegaría a quedar planteado.

Acaso menos encandilado que Meyer con la traducción de los enunciados psicológicos al lenguaje de las ciencias del cerebro –aunque su concepción de la psicología fuera, a la postre, netamente reduccionista–, Hull podía concederse un margen mayor a la hora de interpretar los proyectos de simulación mecánica de procesos psicológicos, o neurológicos. Desde luego –ya se ha dicho–, no esperaba que un modelo mecánico de los reflejos condicionados tuviera que salivar, sino sólo que replicara las "[…] relaciones funcionales entre estímulos y respuestas" (Cordeschi 2002: 109) –seguramente Meyer tampoco habría exigido tanto, ni se habría conformado con tan poco. Plena conformidad a este respecto había entre Hull y Ross, para quien la ya aludida "condición suficiente" que nos autorizaría a atribuir aprendizaje a una máquina –como también recuerda Cordeschi (2002: 247)– "[…] was provided by the functional relations between stimulus and response, without calling into play the specific underlying structure or substratum, whether organic or inorganic".

También Herbert Edgard Coburn (1951, 1952, 1953a, 1953b, *infra*), un ingeniero civil de Minnesota afincado en San Diego, dejaba claro al describir en *The Psychological Review* su "Brain Analogy" que el objetivo que se había fijado era "[…] simular la función, no la estructura, de los sistemas nerviosos orgánicos" (Coburn 1951: 155), pues su proyecto orbitaba en torno a la aspiración de identificar los "[…] principios de los mecanismos inteligentes". La misma idea –simular funciones, no estructuras– se repetiría en el trabajo de Anthony G. Oettinger (1952: 1261), que había articulado un programa capaz de ajustar las respuestas del ordenador a distintos estímulos en función de la aprobación o desaprobación que le comunicara un operador humano –estímulos, respuestas y reforzadores eran siempre números enteros–, y otro que simulaba un proceso de aprendizaje por ensayo y error, ingeniosamente dramatizado bajo la figura del comportamiento de un niño que tiene que ir eligiendo los comercios a los que se dirige para ocuparse de una lista de recados. Mientras Oettinger trabajaba en Cambridge, J. Anthony Deutsch, con sólo veintiséis años, había logrado poner en marcha en el Instituto de Psicología Experimental de Oxford un pequeño aparato –el "maze-runner"– capaz de recorrer laberintos. Acaso con el arrojo que le diera la juventud, anunció que su invento, que luego bautizaría como "[…t]he Insightful Learning Machine" (Deutsch 1955), perfilaba "[…u]na nueva clase de teoría de la conducta" (Deutsch 1953), una teoría psicológica que podía tener otras muchas "realizaciones" o "interpretaciones" (Deutsch 1954: 7), tanto orgánicas como inorgánicas, físicamente dispares. Cada vez

más lejos del espíritu de Meyer, además, Deutsch renunciaba a que las "especulaciones fisiológicas" formaran parte de sus teorías, aun asumiendo que a la larga éstas encontrarían su sustento natural en la neurología; además, al caracterizar su teoría como un "sistema formal", "abstracto", prefiguraba vagamente la vigorosa influencia que el programa formalista de Hilbert para la fundamentación de las matemáticas ejercería –como veremos– sobre la concepción de la mente propia del funcionalismo.

Pero sería K.J.W. Craik (1943), a la sazón *fellow* del Saint John's College de Cambridge, quien con mayor agudeza comenzaría a preguntarse por los principios que sustentan una similitud funcional, como la que podía observarse entre distintas máquinas o entre máquinas y organismos, allí donde no hay similitud estructural. Esos principios venían identificándose sin más con la equivalencia conductual –así, como hemos visto al vuelo, sucede en Ross; también en Oettinger, que asume como criterio el juego de imitación descrito poco antes por Turing (1950). Sin embargo, era precisamente en la elucidación de la idea de similitud funcional donde, más acaso que en la propia analogía entre organismos y máquinas, podía entrever Craik la contribución crucial de los nuevos autómatas a la comprensión de la mente. Al tratar de aislar esos principios, por otra parte, resultaría inexcusable abordar un análisis abstracto de las tareas –en la línea ya entonces emprendida por Turing (1936, 1937):

> It is perhaps better to start with a definite idea as to the kind of tasks a mechanism can accomplish in calculation, and the tasks it would have to accomplish in order to play a part in thought, rather than to draw analogies between the nervous system and some specific mechanism […] and leave the matter there. A telephone exchange may resemble the nervous system […]; but the essential point is the principle underlying the similarity (Craik 1943: 52-53)

Esa lectura fresca, razonablemente saneada de adherencias conductistas o fisicalistas, de lo que podían enseñarnos acerca de los principios que "subyacen a la similitud" entre mentes y máquinas las investigaciones de Turing sobre la noción de autómata tendría que esperar hasta Putnam (1967a, 1967b, *supra*).

Entretanto, este creciente interés en torno a la vida mental de algunas máquinas –por emplear la fórmula felizmente acuñada por Putnam (1967b)– tampoco sería extraño al conductismo. El propio Hull había reflexionado acerca de la posibilidad de construir robots "[…] que dupliquen las capacidades del organismo humano consciente" –así lo recuerda Boring (1950: 673). No en vano, en una conferencia dictada en 1925 en la Universidad de Clark bajo el título "Men or Robots?"[104] William McDougall –quien desde que en 1911 publicara *Body and Mind* se

[104] El neologismo "robot" había sido introducido poco antes, en una obra de teatro de Karel Čapek estrenada en Praga en 1921 –*R.U.R.: Rossumovi univerzální roboti*– que en 1922 cosechó un sonado éxito en Broadway, con Spencer Tracy en el papel de uno de los robots. "Robota", según el *American Heritage Dictionary of the English Language* (Boston: Houghton Mifflin), designaba originalmente, en checo, a los trabajos forzados que los siervos ("rab") debían cumplir en las tierras de sus amos, de modo que el vínculo con la política de organización del trabajo que subraya Shotter (1997, *infra*) es

hallaba embarcado en una ardiente defensa del vitalismo, o, como él prefería decir, del animismo– había aducido que los conductistas erraban radicalmente al dar por supuesta la identidad entre humanos y máquinas[105]. Los conductistas, en realidad, tomaron partido espontáneamente por el mecanicismo en una disputa que los precedía –si bien se había desplegado sobre todo en el ámbito de la biología–, y en cuya otra orilla aún se escuchaba el eco del pensamiento de George E. Stahl (*cf*. *supra*): es a modo de prueba material de que una máquina puede modificar su funcionamiento a la luz de la experiencia sin el concurso de ninguna suerte de protoplasma, entelequia o *élan vital* –y así, al menos en un sentido mínimo, aprender– que Max F. Meyer (1913), en los albores del conductismo, apela al ingenio que había descrito S. Bent Russell en *Psychological Review*. Con su simulación electromecánica de la ley del efecto, John M. Stephens (1929) logró asimismo recabar la atención de Clark L. Hull, quien durante un tiempo se erigiría como heraldo de un *enfoque del robot* llamado algún día a fabricar artefactos ultra-automáticos a los que podríamos sin reparos bautizar como *máquinas psíquicas*.

La intrahistoria de aquellos tiempos se intuye en el diligente relato de Cordeschi (2002: 85): en 1928, un año después de haberla anotado en sus diarios, Hull expone al seminario que cada semana dirigía en la Universidad de Wisconsin-Madison, casi como un reto, la idea de construir un autómata que respondiera al condicionamiento; sólo ha de esperar a la siguiente sesión para encontrarse con tres intentos sobre la mesa, uno de los cuales se presentaría en 1929, en la cuarta conferencia anual de la *Midwestern Psychological Association* en Urbana, Illinois, y poco después en un artículo en *Science* (Hull y Baerstein 1929, *cf*. también Baerstein y Hull 1931). Suma y sigue: Krueger y Hull (1931) pronto describirían un ingenioso circuito eléctrico que –sin que mediara "[…] nous, entelechy, soul, spirit, ego, mind, consciousness, or *Einsicht*"[106] (Krueger y Hull 1931: 267)– parecía sensible no sólo ya al condicionamiento pavloviano como tal, sino también a la extinción, la recuperación espontánea, la irradiación de la respuesta condicionada a estímulos

innegable. Los robots fabricados en la imaginaria factoría Rossum, en todo caso, eran de naturaleza orgánica.

La conferencia de McDougall, que no hace mención expresa a Čapek, fue editada por Murchison (1928), junto con otras impartidas en la misma sede, con firmas tan prominentes como las de John B. Watson, Wolfgang Köhler o Kurt Koffka, entre otros.

[105] Si lo que McDougall atribuía a los conductistas, e impugnaba, era la tesis de que para explicar la conducta es suficiente apelar a un autómata, lo que pensadores afines a la órbita conductista le achacaban a él era la idea de que para explicar la conducta hiciera falta apelar a un fantasma. El trabajo de Meyer (1912) mencionado *supra*, en el que ciertas tesis eliminacionistas aparecen en una de sus formulaciones más tempranas, era de hecho en buena medida una respuesta al libro de McDougall.

[106] El fenómeno que los gestaltistas acostumbraban a describir como *insight*, y que las investigaciones de Wolfgang Köhler (1917) sobre resolución de problemas en chimpancés habían puesto de relieve, sería uno de los blancos explícitos del método de simulación mecánica vehementemente propugnado por Hull, quien –como nos recuerda Cordeschi (2002: 105)– se mostraba convencido de haber proporcionado "[…] a deduction of insight in terms such that it might conceivably be constructed by a clever engineer" (Hull 1935: 231).

semejantes al condicionado, o la posibilidad de generar estímulos condicionados complejos, que dio en llamarse *redintegración* y que Hull consideraba el fundamento de las formas superiores de aprendizaje. Eso sí: incluso una explicación del aprendizaje enteramente purgada de "fuerzas espirituales o sobrenaturales" (Hull 1930: 514) podía permitirse la tesis de que el mundo imprime sobre el organismo una suerte de *representación* de sí mismo, y de que es tal representación, que Hull califica sin ambages de *subjetiva* a la vez que de *funcional*, lo que controla –en un sentido abiertamente causal– la conducta del organismo (Hull 1930: 513). Ahora bien, que esa réplica del entorno sea subjetiva y funcional no es óbice, a ojos de Hull, para que el modo en que queda impresa en el organismo venga dado por determinados circuitos neurofisiológicos[107].

Otros investigadores de una manera u otra ligados al conductismo construyeron asimismo a principios de la década de los cincuenta –como también Leahey (2005: 392) nos recuerda someramente– modelos eléctricos o electromecánicos capaces de remedar los comportamientos que se venían estudiando en los laboratorios conductistas. La idea de una máquina que pudiera aprender a hallar la salida de un laberinto –que Deutsch lograría materializar en 1953– ya había rondado a Hull en 1931, e incluso había sido puesta en práctica por Ross (1933, 1935), que amparaba su investigación bajo el enfoque del propio Hull (*cf*. Cordeschi 2002: 105-106). También Lewis Benjamin Wyckoff, quien en la tesis doctoral que bajo la supervisión del mismísimo Skinner defendiera en 1951 en la Universidad de Indiana había introducido en el análisis conductual el concepto de respuesta de observación, presentaría un modelo electrónico del aprendizaje inspirado en los planteamientos skinnerianos (Wyckoff 1954). Poco después, un psicólogo y psiquiatra de la Universidad de Chicago –aunque formado en Harvard– que buscaba en la noción de sistema las herramientas para forjar una teoría general de la conducta, James G. Miller, alcanzaba ya a entrever "[…] una psicología comparada […] que no trabajase con animales sino con modelos electrónicos" (Miller 1955: 523)[108].

---

[107] También hace notar Cordeschi (2002: 245) que cabría, con Gallistel (1997), rechazar la tesis de que Hull mantuviera una cierta concepción representacional de determinados procesos mediacionales, en la medida en que es del reforzamiento de ciertas conexiones nerviosas de lo que se habla cuando se habla de una representación interna del entorno. Los graves reparos con que Hull acogería la noción de mapa cognitivo propuesta por Tolman (1948) serían entonces un signo de que no procede aribuirle convicciones representacionalistas. Es quizá la inestabilidad entre conclusiones mentalistas y compromisos antimentalistas inherente al conductismo mediacional lo que da aire a la controversia entre Gallistel y Cordeschi, quien por otra parte conviene sin reparos en que para Hull "[…] representations are coded in organisms as neural circuits or patterns in the nervous system, through mechanisms hypothesized by the connectionist and associationist tradition" (Cordeschi 2002: 242).

[108] Anticipándose al mismo tiempo a la apertura interdisciplinar que ha sido característica del cognitivismo, Miller, dicho sea de paso, había colaborado activamente en la formación de un comité encargado de evaluar la viabilidad de esa teoría general de la conducta, para el que se recabaron las contribuciones de psicólogos, psiquiatras, fisiólogos, biólogos, matemáticos, sociólogos, politólogos, antropólogos, economistas e historiadores –*cf*. Duncan (1972: 513). Mucho antes, S.B. Russell (1913) acompañó la presentación de su trabajo con un llamamiento a la colaboración entre "[…] trabajadores

Sin embargo, ni Watson ni Skinner prestaron mayor atención a los robots de Hull, que Boring (1946: 184), pese a su vivo interés en la idea de simulación mecánica, consignaría junto a otras "míseras imitaciones" de la capacidad de aprendizaje de los seres vivos. Acaso desalentado –como mantienen Smith (1986: 358) y Cordeschi (2002: 113)– por la tibia reacción que su proyecto despertaba, el propio Hull iría poco abandonando, a partir de su traslado a Yale en 1934, aquella aproximación sintética al aprendizaje que encarnaban sus máquinas psíquicas, pronto relegadas a la categoría de divertimentos o, como mucho, de herramientas retóricas en la pugna contra el vitalismo. En sus *Principles of Behavior* (1943), donde Hull se refiere ya sólo de pasada a su trabajo en Wisconsin-Madison como el *enfoque del robot*, "[...] the idea of the machine as a tool in theory building and testing" –como acertadamente ha resumido Cordeschi (2002: 114)– "seems to evaporate in the therapeutic function of the machine as a prophylaxis against anthropomorphism".

Quizá debido al fracaso del que había sido el ensayo más vigoroso de dotar al conductismo de una vertiente sintética, todavía Miller, Galanter y Pribram (1960: 51), se ven en la necesidad de argumentar que la simulación computacional es una estrategia explicativa legítima, y, a fuer de hacerlo, comienzan por mofarse de las reticencias de algunos conductistas a una práctica a la que –como fehacientemente había hecho ver Boring (1946)[109]– ellos mismos, de la mano de Hull, habían recurrido a su modo. Que el modelo mecanicista que emplearan fuera por lo general la centralita telefónica, y no la computadora digital, no había de tomarse –era el velado argumento de Miller, Galanter y Pribram– sino como una desafortunada limitación de los empeños explicativos del conductismo: no hay más ni menos rigor mecanicista

---

de diferentes campos de conocimiento" de cara al estudio del sistema nervioso, un área a la que, a su entender, la ingeniería debía contribuir de manera crucial –*cf*. Cordeschi (2002: 67-68).

[109] De hecho, el propio Boring (1946) esboza en las conclusiones de "Mind and Mechanism" una idea que Leahey (2005: 392) ha descrito como "[…] su propia versión de la prueba de Turing": "[…] un robot al que no se le pudiera distinguir de un estudiante sería una prueba extraordinariamente convincente de la naturaleza mecánica del hombre y de la unidad de la ciencia" (Boring 1946: 192). Ya en 1918, discutiendo las observaciones de William James (1907) sobre la "autómata enamorada", Boyd H. Bode había defendido una posición semejante (*cf*. Leahey 2005: 348-349).

Más de un siglo antes, en un desasosegante relato titulado *El hombre de arena*, Ernst T.A. Hoffmann (1816) imaginó que una bellísima autómata llamada Olimpia había sido llevada "[…] de forma encubierta […] a las honestas reuniones de sociedad […] para que ocupara el lugar de una persona de carne y hueso" y que, de hecho, "[…] había acudido a varios 'tés' con éxito". De no ser porque Olimpia era tan hermosa como callada –suspiraba y asentía lánguidamente, quizá como tantas damas del Romanticismo –, se diría que había logrado, en la ficción, superar con creces el juego de imitación. Una muestra rotunda del rechazo que despierta en nosotros algo así reside, como apuntan Bueno y Peirano 2009: 209, en el hecho de que Olimpia fuera tomada por Ernst Jenstch (1906) como paradigma de lo siniestro en su ensayo *Zur Psychologie des Unheimlichen*, que luego Freud (1919) tomaría como punto de partida para sus investigaciones sobre lo siniestro y la angustia de castración. Lo que Jenstch encontraba primordialmente siniestro en Olimpia, sin embargo, era cuanto en ella evocaba a un *Doppelgänger* como los que poblaban las novelas de Jean Paul: sombras fantasmagóricas de una persona que caminan sigilosamente a su lado. Cierta versión de la idea del doble, que tiene ya una remota semilla en *Crátilo* 432*b-c*, desempeñará como veremos un papel capital en la controversia acerca de la eficacia causal de lo mental que vertebra los fundamentos del cognitivismo.

en un modelo que en otro. Ya Edward Tolman, por lo demás, había abogado por abandonar la metáfora de la centralita, que había sugerido en su día William James (1890: 26), y reemplazarla por la idea de una sala de mapas en la que antes de darse autorización a respuesta alguna se desarrolla una escrupulosa consulta de los archivos y protocolos allí almacenados –como se ocupaban de recordarnos Bruner y Goodnow en el prefacio a uno de los textos fundacionales del cognitivismo, *A Study of Thinking* (Bruner, Goodnow y Austin 1956: *vii*[110]). Pero la sala de mapas no era una metáfora tan poderosa como la de la computadora, quizá por no resultar, dado su fuerte carácter homuncular, tan conspicuamente mecanicista. En efecto, en una admirable síntesis de lo que el desarrollo de la computadora supuso para la psicología, en el que destellan tanto la compatibilidad del funcionalismo con el fisicalismo como su neutralidad ontológica, Miller, Galanter y Pribram (1960: 52) escribían que

> […e]l teórico del reflejo ya no es el único psicólogo que puede emplazar a un mecanismo tangible a que torne sus pretensiones en algo más razonable. Hoy el teórico cognitivo también puede convertirse libremente en un materialista integral, si eso es lo que desea.

O bien, como desde el parapeto de una nota a pie de página quedaba punzantemente planteada la cuestión:

> […r]esulta gracioso que tantos psicólogos que abjuran del subjetivismo y del antropomorfismo coloquen sin vacilar centralitas telefónicas dentro de nuestras cabezas. En 1943, por ejemplo, Clark Hull, en sus *Principles of Behavior* […], podía considerar como evidente por sí mismo que el cerebro "actúa como una especie de tablero autómatico de conexión" [Hull 1943: 18, 384]. Sin embargo, el importante adjetivo "automático" es un logro reciente. Los ingenieros de telecomunicaciones que tuvieron que construir y mantener aquellos primitivos tableros de conexión que tanto gustaban a los teóricos del reflejo no se sentían satisfechos con ellos, puesto que requerían de un operador humano que efectuara las conexiones. Por supuesto, dispositivos más perfeccionados acabaron por reemplazar al operador, haciendo que la teoría del reflejo fuera por fin impecable desde el punto de vista científico. Pero en 1892, cuando Karl Pearson escribió *The Grammar of Science*, no tuvo ningún reparo en disponer de un "empleado" que desempeñaba en el cerebro los mismos valiosos servicios que si estuviera en una central telefónica. (Miller, Galanter y Pribram 1960: 51-52)

No está de más preguntarse, entonces, en qué medida Hull se anticipaba a algunas de las posiciones que serían características de los primeros psicólogos cognitivos. Parece razonable acatar el veredicto según el cual la visión de la psicología de Hull, tan discrepante de la de Watson como de la propugnada por Tolman, "[…] terminó por imponerse", mediada la década de 1930, "[…] porque su cuantificación y formalismo lógico pasaban por ser la mejor solución a la crisis de desunión en que estaba sumida la psicología" (Gondra 1992: 17), pero es razonable, también, conceder

[110] La autoría del prefacio y la de la obra difieren porque George A. Austin falleció unos meses antes de que se escribiera aquél, cuando ésta estaba ya entregada al editor.

que el sistema teórico articulado por Hull (1943) distaba mucho de reconocer variables propiamente psicológicas capaces de guiar la conducta, algo a lo que Hull siempre se había mostrado muy reacio. Ahora bien, no es menos cierto que Hull acabaría por dar la razón a Tolman a ese respecto: como ha recordado Mora (1992b: 93), al introducir el concepto de motivación de incentivo, con el que pretendía dar cuenta de los factores cognitivos y motivacionales refractarios al análisis de la conducta bajo el prisma de las nociones de impulso y fuerza del hábit que emergían en los paradigmas de aprendizaje latente y cambio de incentivos, Hull no hacía sino ceder a la insistencia de Tolman respecto a la importancia de la regulación y el control interno de la conducta. No es raro, pues, que tanto Hull como Tolman acabaran siendo blanco de las invectivas –o más bien del desdén– de Skinner (1987). Para el conductista radical, al menos, no había lugar a dudas: tanto Hull como Tolman se habían anticipado a las posiciones cognitivistas, a aquella "restauración cognitiva de la Casa Real de la Mente" (Skinner 1987: 784) en la que Skinner (1987: 783) no veía sino "un encantador adjetivo" pergeñado para obtener, mediante la engañosa promesa de una ciencia de los procesos psicológicos, respaldo económico para líneas de investigación disparatadas y estériles.

La extenuación del computador: contra *capitis defatigatione, mathesis universalis*

La ineluctable presencia en el seno del conductismo de la idea de que una máquina pudiera ser un sujeto psicológico aviva la sospecha de que una concepción de los estados mentales inspirada en el análisis filosófico de la teoría de autómatas habría resultado tal vez algo menos seductora de no haber sido por la fascinación que ha venido ejerciendo en determinados círculos cierta idea mistificada de los trabajos de Alan M. Turing y Alonzo Church, según la cual estos habrían convergido en la demostración de que para cualquier función que la mente pueda computar existe un autómata que puede también computarla. Pero, como entre otros ha denunciado Copeland (2002), los argumentos desplegados por Turing y Church distan de establecer tal cosa; en realidad, ni siquiera lo pretenden.

Cuando Turing (1936, 1937, *supra*) describió por primera vez lo que acabaríamos conociendo como máquinas de Turing, su propósito era aclarar las nociones de tarea computable, procedimiento efectivo, o mecánico, y algoritmo, a todas luces estrechamente ligadas entre sí. Resulta claro de la lectura de sus trabajos que el modelo que opera en el pensamiento de Turing entonces es el del oficinista o el burócrata[111], enfrascado, como plásticamente nos recuerda Copeland (2002), en "[…] a certain human activity, the tedious one of *numerical computation*, which until the advent of automatic computing machines was the occupation of many thousands

[111] El estudio del papel de la metáfora burocrática en el cognitivismo, y en particular en la conformación de su sustrato ideológico, se abordará brevemente *infra* al hilo del trabajo de Shotter (1997).

of people in business, government, and research establishments"[112]. El propio Turing (1948: 9) había insistido en que "[…a] man provided with paper, pencil, and rubber, and subject to strict discipline, is in effect a universal machine", y usaba una y otra vez el vocablo "computer" y otros de su campo semántico para referirse no a máquinas sino a trabajadores (*cf*. por ejemplo Turing 1947: 116, Turing 1947: 120, Turing 1950a: 56, *infra*; el recuento se debe a Copeland 2002). Lo que interesaba a Turing era la naturaleza de las tareas que puede desempeñar "[…] a human operator working in a disciplined but unintelligent manner" (Turing 1950b: 1), que pueden abordarse mediante "[…] human clerical labor, working to fixed rules, and without understanding" (Turing 1946: 38-39). Incluso la descripción primera de las máquinas de computación lógica se abre con la misma idea: "We may compare a man in the process of computing […] to a machine" (Turing 1936: 231). La cuestión, como también registra Copeland, no pasó inadvertida para Wittgenstein (1946-1949/1980: §1096), quien con su inconfundible laconismo apuntó "Turing's 'Machines'. These machines are humans who calculate".

"Computadores" o "calculadores" habían venido siendo desde antiguo los monjes dedicados al cómputo eclesiástico –la determinación de la fecha de la Pascua de Resurrección de acuerdo con los procedimientos fijados en el año 325 por el Concilio de Nicea y las rectificaciones introducidas doscientos años después por Dionisio el Exiguo–; luego lo serían también quienes se encargaban de la confección de almanaques, portulanos y cartas náuticas, a la teneduría de libros contables, o a la elaboración de registros astronómicos. Su índole rutinaria era, junto a la frecuencia con que se deslizaban errores en el cálculo, lo más característico de las tareas de los computadores: ya en un tratado renacentista dedicado a la instrucción militar como es *Teoría y Práctica de Fortificación*, del capitán Cristóbal de Rojas –que fuera ayudante de Juan de Herrera–, se anota que la comprensión de determinados procedimientos matemáticos basta para los intereses del soldado o el ingeniero castrense, "[…] si bien la tal inteligencia será mecánica" (Rojas 1598: fol. 1 *apud* Guijarro y González 2010: 83). Sería quizá más que ninguna otra cosa –como sostienen precisamente Guijarro y González 2010: 18)– la necesidad de mitigar la plaga de pequeños errores que enturbiaba los múltiples registros tabulares que cada vez más requería la organización estatal, así como de acortar el tiempo requerido para su elaboración, lo que llevaría a Charles Babbage al empeño de construir sus dos máquinas de diferencias, primero, y su máquina analítica, después –a las que Turing (1950a: 59, *infra*) se refiere expresamente.

Las condiciones de trabajo de los computadores, con todo, pronto empezarían también a ser una preocupación; su *capitis defatigatione*, a la que aluden Guijarro y González (2010: 114), era sin duda la más patente. El propio Babbage mencionaría en una carta al eminente químico Sir Humphry Davy, entre los motivos que le hicieron

[112] Al espectador de *El apartamento*, de Billy Wilder (1960), tal vez le vengan a la memoria los planos de la oficina de Consolidate Life donde C.C. "Bud" Baxter (Jack Lemmon), uno entre exactamente 31.259 empleados, trataba de medrar prestando su apartamento para las aventuras extramatrimoniales de sus superiores.

concebir sus máquinas de cálculo, "[…] el intolerable esfuerzo y la fatigosa monotonía de una continua repetición de cálculos aritméticos similares" (Babbage 1822: 44 *apud* Guijarro y González 2010: 325), que él había sufrido en carne propia mientras colaboró en las tareas de la *Analytical Society* –una institución dedicada a promover la sustitución de la notación newtoniana por la leibniziana para el cálculo infinitesimal, de la que fue cofundador en sus tiempos de estudiante. Ya Bowden (1953)[113] relata cómo la ardua labor de los computadores inspiró los audaces proyectos de Babbage:

> In 1812 he was sitting in his rooms in the Analytical Society looking at a table of logarithms, which he knew to be full of mistakes, when the idea occurred to him of computing all tabular functions by machinery. The French government had produced several tables by a new method. Three or four of their mathematicians decided how to compute the tables, half a dozen more broke down the operations into simple stages, and the work itself, which was restricted to addition and subtraction, was done by eighty computers who knew only these two arithmetical processes. Here, for the first time, mass production was applied to arithmetic, and Babbage was seized by the idea that the labours of the unskilled computers could be taken over completely by machinery which would be quicker and more reliable.

Años después, cuando al cabo de los viajes que emprendió para familiarizarse con el tejido industrial redactara *On the Economy of Machinery and Manufacturers* (1832), no dejó de reclamar una mayor participación de los empleados en la organización de la fábrica, así como la limitación del trabajo infantil. Sin embargo, los errores de cálculo o de copia proporcionaban a Babbage una retórica mucho más vigorosa en su incansable esfuerzo por recaudar los copiosos fondos que precisaba para la construcción de sus ingenios: mientras esos errores siguieran infectando tablas astronómicas, náuticas o fiscales, seguirían provocando quebrantos acaso más cuantiosos que la inversión que Babbage reclamaba, e incluso pérdidas de vidas humanas. En la misma carta a Davy aseguraba Babbage (1822: 45 *apud* Guijarro y González 2010: 237) que la lista de errores del *Almanaque Náutico* elaborado por el *Board of Longitude*, que se había tomado el trabajo de examinar, era, por increíble que pudiera parecer, más larga que la lista original de datos –seis años después, el Parlamento aboliría el *Board*.

No muy lejos parece quedar, en todo caso, la idea de los computadores humanos que Turing tiene en mente de lo que el *Manifiesto del Partido Comunista* había denunciado, casi un siglo atrás, en los años en que Babbage trabajaba en su invento: que "[… al obrero] sólo se le exigen las operaciones más sencillas, más monótonas, de más fácil aprendizaje" (Marx y Engels 1848: 20) –o, como rezaba la

[113] En una crestomatía de ensayos sobre los nuevos "cerebros electrónicos" a la que el propio Turing contribuyó con una disertación sobre la importancia teórica y técnica de la programación de máquinas capaces de jugar cuyas páginas se abrían con menciones del falso autómata ajedrecista de Wolfgang von Kempelen –"el Turco", *cf. infra*– y del verdadero, aunque mucho más humilde, autómata ajedrecista que Leonardo Torres Quevedo había presentado en París en 1914, que se limitaba a dar jaque a un solitario rey sin más ayuda que la de una torre y –claro– la de su propio rey.

primera traducción al castellano del Manifiesto, que en 1872 publicara en Madrid *La Emancipación*, "[…] una operación fatigosa, monótona y puramente mecánica" (Marx y Engels 1848: 246) –el propio Marx, de hecho, cita repetidamente la descripción de la vida fabril trazada por Babbage (*cf*. Blaug, ed. 1991). La Universidad de Cambridge era en los años treinta un hervidero ideológico, y el joven Turing tal vez habría acabado incorporándose a las filas comunistas de no haber sido porque su indócil talante pronto lo hizo sentirse en casa en King's Collage, donde John Maynard Keynes había creado un espacio de librepensamiento desvinculado de toda moral convencional –incluida, claro, la soviética. En mayo de 1933, tras dos años como estudiante, Turing escribía a su madre para, además de agradecerle el envío de unos calcetines, contarle que estaba pensando en "[…] going to Russia some time in vac[ation] but have not yet quite made up my mind. [/] I have joined an organisation called the 'Anti-War Council.' Politically rather communist" (Hodges 1983: 71). Aunque el compromiso ideológico de Turing se desvanecería pronto, y nunca mostró interés por las pretensiones marxistas de proporcionar una explicación científica de los hechos históricos[114], no es aventurado distinguir la huella de aquellos años en su preocupación por los aspectos más rutinarios y mecánicos del trabajo.

Pues bien, el punto de partida del trabajo de Turing es la idea de que un procedimiento efectivo –mecánico, algorítmico– para llevar a cabo una tarea puede entenderse intuitivamente como aquel que consta de un número finito de instrucciones inequívocas (y ellas mismas finitas) que podría seguir un operario sin otra ayuda que papel y lápiz –y, en particular, sin que le fuera exigido ningún ingenio especial[115]–, y que conduce con toda seguridad a la resolución de la tarea. Las tareas que pueden abordarse con éxito mediante procedimientos efectivos son tareas computables. Lo que Turing postuló –lo que bien podríamos denominar "tesis de Turing"– es que para toda tarea que resulte ser, en este sentido, una tarea computable, o "puramente mecánica" (Turing 1948: 7), hay una "computadora lógica mecánica" –una máquina de Turing– capaz de resolverla. Afirmar, así pues, que una tarea es computable sería tanto como afirmar que puede resolverla una máquina de Turing: dicho de otro modo, "computable" y "Turing-computable" serían expresiones equivalentes.

Miríadas de trabajadores cuyos talentos tendrían mejor empleo en otras tareas, entonces, podrían pronto quedar reemplazados por computadoras: sólo sería preciso diseñar y producir una máquina dedicada a cada una de las tareas que se nos presentaran. En la visión de Turing, sin embargo, se vislumbraba un escenario sustancialmente más ambicioso, alumbrado por una única máquina capaz de desempeñar toda tarea que pudiera llevar a cabo cualquiera de las máquinas que él mismo había descrito –en definitiva, una máquina universal:

---

[114] El capítulo 2 de Hodges (1983) contiene algunas pistas sobre el pensamiento político de Turing en sus años de estudiante universitario, y en particular sobre sus relaciones con los economistas del King's, como Arthur Pigou o el propio Keynes.

[115] Ninguna suerte de "espabilamiento", dicen Mosterín y Torretti (2002: 354) en una derivación neológica de aire arcaizante que resulta, precisamente, de lo más ingenioso.

> The importance of the universal machine is clear. We do not need to have an infinity of different machines doing different jobs. A single one will suffice. The engineering problem of producing various machines for various jobs is replaced by the office work of 'programming'[116] the universal machine to do these jobs. (Turing 1948: 7)

Una máquina universal de Turing puede, por tanto, remedar el funcionamiento de cualquier máquina de Turing. Como escuetamente queda expresado en Barker-Plummer (2004), "[…w]hen started on a tape containing the encoding of another Turing machine[117], call it *T*, followed by the input to *T*, a U[niversal] T[uring] M[achine] produces the same result as *T* would when started on that input", donde bajo la noción de idéntico resultado se obvian diferencias en cuanto al tiempo de ejecución o al número de transiciones.

Poco antes que Turing, Alonzo Church había alcanzado conclusiones virtualmente idénticas empleando para reconstruir del concepto de algoritmo, en lugar de las máquinas abstractas pensadas por Turing, un aparato lógico bautizado como cálculo lambda, que acababan de construir él mismo (Church 1932, 1936a, 1936b) y Stephen Kleene (1935, 1936). Así, la clase de las funciones intuitivamente computables sobre números enteros positivos resultaba equivalente a la clase de funciones que el cálculo lambda ofrecía recursos para definir: "computable" sería tanto como "*lambda-definible*". El mismo Turing pronto establecería –en un apéndice añadido a su trabajo de 1936 durante el proceso de revisión, cuando tuvo conocimiento de los resultados de Church– que su propia noción de computabilidad, de aplicación más general, y la noción de *lambda-definibilidad* de Church eran intercambiables siempre que nos ciñéramos al ámbito de aplicación del trabajo de Church, más restringido. También lo eran, por tanto, con los mismos matices, la tesis de Turing –un procedimiento es computable si y sólo si es *Turing-computable*– y la tesis de Church –una función de enteros positivos es efectivamente calculable si y

[116] Que Turing tenga que escribir entrecomillado el verbo "programar" es un delicado signo de su condición de pionero.

[117] La idea de *codificación de una máquina de Turing* requiere aclaración. Siguiendo de nuevo a Barker-Plummer (2004), cada una de las cuádruplas {$Estado_i$, *Símbolo*, $Estado_f$, *Acción*} que forman la tabla de máquina de una máquina de Turing se puede codificar como cuatro secuencias de unos separadas por un cero: la primera secuencia, $Estado_i$, representa el estado inicial de la máquina en esa regla de transición, de forma que si el número asignado a dicho estado es *n*, la secuencia contiene *n*+1 unos; la segunda secuencia, *Símbolo*, contiene un uno si el símbolo es *0* y dos si es *1*; la tercera secuencia, $Estado_f$, representa al estado final especificado en la regla de transición mediante el mismo código usado en la primera secuencia; la última secuencia, *Acción*, contiene un uno si la acción fijada es reemplazar el símbolo leído con un *0*, dos unos si se trata de reemplazarlo con un *1*, tres uno cuando la acción sea un desplazamiento a la izquierda, y cuatro para el desplazamiento a la derecha. Si cada grupo de cuatro secuencias, que representa la cuádrupla de una regla de transición, se aisla encabezándola y finalizándola con un cero –puesto que el cero no ha sido utilizado en la codificación–, y la tabla de máquina completa se aisla encabezándola y finalizándola con dos ceros seguidos, habremos obtenido una larga secuencia identificable de unos y ceros que representa la totalidad de la tabla de máquina de una máquina de Turing, y que nos proporciona el *número de serie* de la máquina, susceptible de interpretación como un número binario.

sólo si es *lambda-definible*–; de ahí que hiciera fortuna el giro "tesis de Church-Turing"[118]. Los motivos por los que suele preferirse la caracterización del concepto intuitivo de procedimiento efectivo ensayada por Turing a otras formalmente equivalentes, como la de Church, es que resulta intuitivamente más convincente; los motivos por los que resulta intuitivamente más convincente evocan vivamente la reflexiones germinales de Putnam (1967a, 1967b) sobre cuánta luz podía arrojar sobre nuestra comprensión de lo mental el estudio de los autómatas abstractos. En Nagornyi y Marchenkov (2001), por ejemplo, dicha evocación cobra la forma de un recorrido en sentido contrario por el mismo camino: a su juicio, si la tesis de Turing resulta más persuasiva que la tesis de Church o que cualquier otra formulación equivalente, es porque

> […] by carrying out computations according to a selected plan, the mathematician acts in a similar way to a Turing machine: in considering some position in his writings and being in a certain "state of mind", he makes the necessary alterations in his writings, is inspired by a new "state of mind", and goes on to contemplate further writing. The fact that he completes more complicated steps than a Turing machine seems not principally significant.

Pero como con minuciosidad implacable ha mostrado Copeland (2002), la tesis de Church-Turing –o tal vez una mudable mezcolanza de la tesis de Church-Turing y la noción de máquina universal– ha sido insistentemente malinterpretada en el seno del debate contemporáneo sobre la explicación psicológica y la naturaleza de lo mental. La semilla de la confusión tal vez se encuentre, como ha sugerido Piccinini (2007), en un desafortunado argumento de John von Neumann (1951), uno de los creadores de

---

[118] Además, el concepto de función lambda-definible resulta equivalente también, tal como se ocuparon de establecer Church (1936a) y Kleene (1936), al de función recursiva, formulado por Herbrand (1932) y Gödel (1934). En casi cualquier manual introductorio a la teoría de la computación puede encontrarse el relato de cómo los intentos posteriores de reconstruir la noción intuitiva de procedimiento efectivo –como las máquinas de Post (1943, 1946), los algoritmos de Markov (1960), etc. – también han resultado ser equivalentes al análisis de Turing, lo que ha asentado la convicción de que la tesis de Church-Turing, aunque la vaguedad inherente a dicha noción intuitiva la haga indemostrable, es verdadera.

Seguramente no está de más hacer notar, en relación con la crítica chomskiana de los modelos lingüísticos basados en cadenas de Markov (*cf*. Chomsky 1956, 1957, *supra*), que los algoritmos normales de Markov no son cadenas o procesos de Markov. La noción de algoritmo normal desarrollada por Markov suele ser considerada, junto con la de función recursiva y la de máquina de Turing, como el análisis más perspicuo del concepto intuitivo de algoritmo (*cf*. por ejemplo Nagornyi y Marchenkov 2001). Ahora bien, en la caracterización de un algoritmo normal de Markov no interviene en absoluto la propiedad de Markov, que es definitioria de los procesos markovianos; de hecho, comoquiera que cada paso del algoritmo depende de una lista finita y ordenada de fórmulas de sustitución, denominada *esquema* del algoritmo –y no sólo, estocásticamente, del resultado del paso anterior–, se verifica que un algoritmo normal de Markov es un proceso no markoviano. Afirmar, pues, que la equivalencia entre la noción de máquina de Turing y la de algoritmo normal de Markov muestra que las críticas de Chomsky a los modelos markovianos en lingüística –o su célebre jerarquía de gramáticas, *cf*. Chomsky 1956, *supra*– eran erradas constituiría una flagrante *equivocatio*.

la arquitectura básica que comparten la mayoría de los ordenadores modernos. En un trabajo presentado en 1948 al simposio de la Fundación Hixon en el Instituto de Tecnología de California –donde también disertó Karl Lashley (1951, *supra*) sobre el problema del orden serial de la conducta–, von Neumann trataba de aquilatar el valor de las recientes contribuciones de McCulloch y Pitts (1943), también presentes en el encuentro, y que se cuentan entre los primeros intentos, casi un cuarto de siglo antes de Putnam (1967a, 1967b), de usar los avances en teoría de la computación para iluminar el estudio de la mente. Agrupadas en redes, las neuronas artificiales –*i.e.*, representaciones matemáticas del funcionamiento de una neurona– que McCulloch y Pitts describían en su "Cálculo lógico de las ideas inmanentes en el sistema nervioso" (1943) sólo desmerecían de la capacidad computacional de una máquina de Turing en la medida en que carecían de dispositivos equivalentes a la cinta infinita y el cabezal de lectura-escritura. Ese resultado –pensaban McCulloch y Pitts– ofrecía un vigoroso respaldo a la tesis de Church-Turing: puesto que la noción intuitiva de procedimiento efectivo –suponían– tiene que ver en el fondo con lo que el cerebro humano –o al menos el cerebro de un oficinista– puede computar, y puesto que la capacidad computacional de las redes neuronales que, *ex hypothesi*, conforman el cerebro humano no excede la de las máquinas de Turing (o, si se prefiere, la del cálculo lambda), podemos concluir que el análisis de la noción intuitiva de procedimiento efectivo propuesto por Turing (o Church) queda amparado por la modelización matemática del funcionamiento del sistema nervioso. Acaso entusiasmado por las perspectivas de confluencia entre la teoría de la computación y las ciencias del cerebro, von Neumann trató de encontrar en la tesis de Church-Turing un apoyo a la idea de que el funcionamiento del cerebro humano se basa en operaciones de naturaleza computacional, invirtiendo –como hace notar Piccinini (2007: 98), quien recoge el mismo pasaje– el orden de justificación ensayado por McCulloch y Pitts:

> The McCulloch–Pitts result [...] proves that anything that can be exhaustively and unambiguously described, anything that can be completely and unambiguously put into words, is *ipso facto* realizable by a suitable finite neural network. Since the converse statement is obvious, we can therefore say that there is no difference between the possibility of describing a real or imagined mode of behavior completely and unambiguously in words, and the possibility of realizing it by a finite formal neural network. The two concepts are coextensive. A difficulty of principle embodying any mode of behavior in such a network can exist only if we are also unable to describe that behavior completely (von Neumann 1951: 22–23).

El propio Piccinini (2007: 99) señala que las palabras de von Neumann constituyen probablemente la expresión primigenia de la falacia de la tesis de Church-Turing, tal como ha sido denunciada por Copeland (2002), que no es sino una interpretación viciada de la tesis de Church-Turing: la de que la tesis de Church-Turing demuestra o entraña que la mente humana es una máquina de Turing. Después, como quien dice, una cosa llevó a la otra, y una laxa lectura de la tesis de Church-Turing según la

cual hay una máquina de Turing capaz de computar cualquier procedimiento claramente definido fue abriendo paso a otra, aún más laxa, según la cual hay una máquina de Turing capaz de simular cualquier fenómeno susceptible de descripción científica[119] –o a menudo, informalmente, cualquier fenómeno, punto.

En las primeras líneas de la propuesta formal que dio pie a la *Summer Research Conference on Artificial Intelligence* de Dartmouth, redactada por John McCarthy, Marvin Minsky, Nathan Rochester y Claude Shannon (1955), se invocaba ya a la "conjetura" de que "[…] every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it". No tardó mucho en quedar olvidado que aquello era, exactamente, una conjetura. Así, Dennett (1978b: *xviii*) nos advierte de que toda tarea que podamos describir claramente como una cadena de pasos simples "[…] can be performed by a very simple computer, a universal Turing machine, the universal recipe-follower"; Gregory (1987: 784), en una obra de consulta, presenta como un hecho cierto, demostrado por el propio Turing, que una máquina de Turing "[…] can specify the steps required for the solution of any problem that can be solved by instructions, explicitly stated rules, or procedures"; Churchland y Churchland (1990: 26) se muestran convencidos de que cualquier ordenador digital, provisto de tiempo, memoria y el programa adecuado, "[…] can compute any rule-governed input-output function […; t]hat is, it can display any systematic pattern of responses to the environment whatsoever"; Dennett (1991a: 215) insiste en que Turing habría demostrado que "[…] his Universal Turing Machine can compute any function that any computer, with any architecture, can compute". Paralelamente, Churchland y Churchland (1983: 6) nos aseguran que si asumimos –lo cual consideran que podemos asumir con garantías– que las operaciones del sistema mente-cerebro son computables, entonces la tesis de Church-Turing nos muestra que "[…] it can in principle be simulated by a computer"; Johnson-Laird (1987: 252), poco después, declara que si asumimos que la consciencia es explicable científicamente y aceptamos la propuesta de que los estados psicológicos se identifican por sus relaciones funcionales –es decir, el funcionalismo–, entonces, dada la tesis de Church-Turing, debemos concluir "[…] la consciencia es un proceso computacional"; Guttenplan (1994: 595), también en un texto de referencia, da por cierto que basta con que las relaciones entre aferencias y eferencias del sistema nervioso muestren un patrón al que pueda darse alguna descripción matemática para que podamos contar con que

[119] La tesis de que una máquina de Turing puede computar todo lo que pueda computar cualquier máquina con instrucciones y datos finitos es denominada por Copeland "tesis M", a la de que todo proceso que pueda ser descrito matemáticamente –o explicado científicamente– puede ser simulado por una máquina de Turing le dará el nombre de "tesis S". De la tesis S se seguiría que existe necesariamente una máquina de Turing capaz de simular el funcionamiento del cerebro, o el de la mente, siempre y cuando asumamos que existe una descripción científica de dicho funcionamiento. Pero ni la tesis M ni la tesis S, desde luego, se derivan de la tesis de Church-Turing.

Todas las referencias del párrafo siguiente provienen de Copeland (2002), que se ha ocupado de espigar estas y otras muchas muestras de involuntaria tergiversación de la tesis de Church-Turing.

"[…] some specific version of a Turing machine will be able to mimic them"; incluso un destacado crítico de la concepción cognitivista de la mente, que, como es el caso de John R. Searle, se ha empeñado en desenmascarar lo que entiende como extravíos de la metáfora computacional, parece preso de lo que Smolensky (1988: 3) había llamado "la interpretación fuerte de la tesis de Church" cuando asevera que "[…] given Church's thesis that anything that can be given a precise enough characterization as a set of steps can be simulated on a digital computer, it follows trivially that the question [whether the operations of the brain can be simulated on a digital computer] has an affirmative answer" (Searle 1992: 200).

La distraída concesión de Searle se encardina, en efecto, en lo que él mismo bautiza como "el relato primigenio" de las razones por las que "[…] el cognitivismo ha resultado intuitivamente atractivo" (Searle 1992: 202), razones a cuya impugnación se dispone a renglón seguido. El planteamiento de ese relato se articula sobre la tesis de Church-Turing y lo que Searle denomina indistintamente "el teorema" o "la tesis" de Turing, que establece la existencia de la máquina universal. El siguiente paso es la idea de que el cerebro sea tal máquina universal; el origen de esa idea sería, según el análisis informal de Searle, un razonamiento más o menos así:

> It is clear that at least some human mental abilities are algorithmic. For example, I can consciously do long division by going through the steps of an algorithm for solving long division problems. It is furthermore a consequence of the Church-Turing thesis and Turing's theorem that anything a human can do algorithmically can be done on a Universal Turing Machine. I can implement, for example, the very same algorithm that I use for long division on a digital computer. In such a case, as described by Turing (1950[a]), both I, the human computer, and the mechanical computer are implementing the same algorithm. I am doing it consciously, the mechanical computer nonconsciously.
>
> Now it seems reasonable to suppose there might also be a whole lot of mental processes going on in my brain nonconsciously which are also computational. And if so, we could find out how the brain works by simulating these very processes on a digital computer. Just as we got a computer simulation of the processes for doing long division, so we could get a computer simulation of the processes for understanding language, visual perception, categorization, etc. (Searle 1992: 202-203)

Con esto, apunta Searle, "[…] tenemos un programa de investigación bien definido". *Voilà*:

> We try to discover the programs being implemented in the brain by programming computers to implement the same programs. We do this in turn by getting the mechanical computer to match the performance of the human computer (i.e. to pass the Turing Test) and then getting the psychologists to look for evidence that the internal processes are the same in the two types of computer. (Searle 1992: 203)

Probablemente otra de las semillas de la mala interpretación de la tesis de Church-Turing resida en que "Computing Machinery and Intelligence" (Turing 1950a) es un trabajo mucho más asequible, y mucho más leido que "On Computable Numbers, With an Application to the *Entscheidungsproblem*" (Turing 1936, 1937). En el artículo

de 1950, como es bien sabido, Turing nos invita a dejar de preguntarnos si "[…] pueden pensar las máquinas" –una pregunta que es, a su juicio, "[…] demasiado insignificante para que amerite discusión" (Turing 1950a: 54)– y a abordar en cambio la cuestión de si la conducta de una computadora digital en un determinado examen de sus capacidades podría llegar a resultar indiscernible de la de un sujeto humano: el examen es, naturalmente, el célebre "juego de imitación" (Turing 1950a: 53), o "test de Turing", en el que un juez independiente que mantiene una conversación tipográfica con un humano y con una máquina debe discriminar cuál es cuál. A la hora de explicar qué es una computadora digital, Turing es tan cuidadoso como había sido en su trabajo de 1936 en lo que atañe a la relación de las capacidades humanas con las de las computadoras digitales:

> La idea detrás de las computadoras digitales puede explicarse diciendo que se trata de máquinas cuyo objetivo es realizar cualquier operación que pueda realizar una computadora humana. Esta computadora humana supuestamente sigue reglas fijas y carece de la autoridad para desviarse de ellas en el más mínimo detalle. (Turing 1950a: 56)

Se trata, pues, del mismo afanado oficinista al que ya habíamos encontrado en Turing (1936, 1947, 1948, 1950b, *supra*), y nada indica por el momento que las computadoras digitales puedan *en principio* remedar cualquier operación que pueda realizar un ser humano y que desborde los restringidos límites de la estricta rutina. Pero como los propósitos del artículo de Turing son fundamentalmente polémicos, esa contención inicial pronto queda atrás, y el argumento se demora en la desestimación de posibles alegaciones en el sentido de que ninguna computadora podría superar el juego de imitación. Con desigual paciencia, Turing trata de ir desarmando cada una de las muchas virtudes humanas que –se aduce– estarían vedadas a las máquinas[120]:

> La capacidad de ser amable, ingenioso, hermoso, amistoso, de tener iniciativa, sentido del humor, de distinguir lo bueno de lo malo, de cometer errores, de enamorarse, de disfrutar las fresas con crema, de lograr que alguien se enamore de ella, de aprender de la experiencia, de usar palabras correctamente, de ser sujeto del propio pensamiento, de tener la misma diversidad de comportamientos que el hombre y de hacer algo en verdad novedoso. (Turing 1950a: 67).

El esfuerzo de Turing por señalar, alternativamente, que acostumbramos a ver agigantadas nuestras bondades en estos ámbitos y a menoscabar de antemano las

[120] Y que él considera facetas de una cuestión más general, la de si una máquina podría albergar estados conscientes (Turing 1950a: 70), que trata de neutralizar subordinándola, a su vez, a otra aún más general, pero de índole epistemológica: cómo podríamos saber si alberga consciencia cualquier ser humano distinto de uno mismo (Turing 1950a: 66).

que las máquinas pudieran mostrar[121] –o bien, en algunos casos, que la capacidad en cuestión es irrelevante de cara a la pregunta por la inteligencia–, acaba inevitablemente dejando la viva impresión de que Turing está defendiendo que toda capacidad humana es susceptible de ser reproducida con toda fidelidad por una computadora digital. Incluso cuando se enfrenta a objeciones enraizadas en su propio trabajo –"resultados de lógica matemática", entre los que se cuentan, junto con los de Gödel (1931), Kleene (19335), o Church (1936), los cosechados por Rosser y el propio Turing (1937), "[…] que pueden utilizarse para demostrar que hay limitaciones al potencial de las máquinas de estado discreto" (Turing 1950a: 64), como son las computadores digitales, la estrategia de Turing discurre por los mismos cauces:

> […] aun cuando se ha determinado que existen limitaciones al poder de cualquier máquina particular, sólo se ha afirmado, sin ningún tipo de comprobación, que ninguna de estas limitaciones se aplica al intelecto humano. No creo, sin embargo, que pueda descartarse tan a la ligera este punto de vista. (Turing 1950a: 65)

Pese a todo, lo cierto es que en ningún momento llega Turing a aseverar que una computadora digital sea capaz de remedar íntegramente el intelecto humano, y menos que tal cosa haya quedado establecida o constituya una cuestión de principio. De hecho, cuando justo antes de afrontar la controversia Turing se dispone a sincerarse y dejar clara su posición, se limita a la predicción de que:

> […] aproximadamente dentro de 50 años será posible programar computadoras […] que tomen parte tan bien en el juego de imitación que el examinador promedio no tenga más del 70% de probabilidad para lograr la identificación correcta luego de cinco minutos de preguntas. (Turing 1950a: 62)

Pero que Turing creyera o no que la totalidad de las destrezas cognitivas humanas consistan en procedimientos computables –procedimientos que, en último término, pudiera reproducir la máquina universal que él mismo había concebido– es, después de todo, poco relevante: la clave, sea como sea, es que ni esa conclusión ni su negación es a lo que apuntan los resultados conocidos como tesis de Church, o tesis de Turing. Una lectura descuidada de Turing 1950a, sin embargo, bien puede instigar precisamente esa inferencia.

El caso, para bien o para mal, es que la razón asiste a Copeland cuando sobriamente nos recuerda que lo único que afirma la tesis de Church-Turing es que cierta clase de procedimientos –aquellos que, según el análisis de Turing, intuitivamente calificaríamos como procedimientos mecánicos, efectivos– es idéntica a la clase de los procedimientos que puede llevar a cabo alguna máquina de Turing: ni una palabra sobre la posibilidad de que determinados sistemas –organismos o

[121] Una táctica, por cierto, que no puede menos que evocar a la que con análogas miras ensayara mucho antes Darwin (1871: *passim*) en su polémica con Alfred Wallace a cuenta del origen de las facultades psicológicas distintivas del ser humano.

máquinas– puedan llevar a cabo procedimientos que desborden el estrecho ámbito de lo que un operario humano, sin más ayuda que lápiz y papel, sería capaz de hacer. En particular,

> [...t]he Church-Turing thesis does not entail that the brain (or the mind, or consciousness) can be modelled by a Turing machine program, not even in conjunction with the belief that the brain (or mind, etc.) is scientifically explicable, or exhibits a systematic pattern of responses to the environment, or is "rule-governed" (etc.). (Copeland 2002)

Más prudente parece, así pues, la posición de Bechtel (1988: 154-155), para quien "[…] en la medida en que los procesos llevados a cabo por la mente son efectivamente computables, hay una Máquina de Turing que será conductistamente [*sic*] equivalente a la mente"[122]. El giro inicial "en la medida en que" recaba los matices que preocupan a Copeland, pero no es aventurado sugerir que cierta tendencia a pasar por alto esos matices ha contribuido espuriamente al auge del cognitivismo de inspiración funcionalista. Eso, por supuesto, no refuta el funcionalismo, pero ayuda a entender sus diversos itinerarios.

De hecho, un somero repaso de las condiciones que el propio Turing exigía para que un determinado procedimiento fuese considerado, en el sentido intuitivo, un procedimento mecánico –*ergo*, de acuerdo con la tesis de Turing, Turing-computable– basta para despertar la sospecha de que algo ha fallado cuando nuestra lectura de Turing parece conducir inexorablemente a la idea de que el funcionamiento del cerebro o de la mente humana es íntegramente susceptible de ser reproducido por una máquina de Turing. Un procedimiento dado constituye un algoritmo –es de naturaleza efectiva, o mecánica– cuando, recuérdese, desemboca más pronto o más tarde en la solución de la tarea por medio de la simple aplicación de un conjunto finito de instrucciones claras y precisas (cada una de las cuales es también finita), instrucciones que un trabajador disciplinado, lapicero en mano, puede poner en práctica sin necesidad de comprensión, ingenio, o inteligencia. Las computadoras lógicas mecánicas de Turing fueron ideadas precisamente con el propósito de reproducir un conjunto limitado de capacidades humanas, las que pueden desplegarse sin comprensión, ingenio o inteligencia. Que las capacidades de dichas computadoras –las máquinas de Turing– coincidan con ese conjunto de capacidades humanas es –qué duda cabe– un logro admirable del análisis de Turing; que las capacidades de las máquinas de Turing dieran en agotar la totalidad de las capacidades humanas, por mucha comprensión, ingenio, inteligencia –sensibilidad, consciencia, etc.– que medie en ellas, sería verdaderamente notable; que tal cosa pudiera quedar *demostrada sin ningún argumento adicional* sería ya de todo punto asombroso –máxime cuando ni siquiera es estrictamente demostrable la

[122] A pesar de que se le atribuye una genealogía algo confusa: a ojos de Bechtel, tal conclusión se deriva de conjugar la Tesis de Church, según la cual todo proceso efectivamente computable sería computable mediante un procedimiento recursivo, y la demostración de que la Máquina de Turing es una máquina universal, en el sentido de que es capaz de computar cualquier procedimiento recursivo.

coextensividad de la clase de procedimientos Turing-computables y la de la clase de procedimientos intuitivamente mecánicos, dada la vaguedad de la definición de esta última. Sin embargo, la interpretación mistificada del trabajo de Turing que ha dado alas a la concepción funcionalista de lo mental sostiene precisamente eso.

El "relato primigenio" de la fascinación cognitivista por las máquinas de Turing esbozado por Searle (1992, *supra*), en realidad, proporciona una armazón para ese argumento: de la misma manera que una máquina de Turing que realiza operaciones aritméticas ejecuta inconscientemente el mismo algoritmo que uno puede ejecutar conscientemente, cabe pensar que el cerebro ejecuta inconscientemente determinados algoritmos cada vez que percibimos o pensamos. Es decir, que los procesos psicológicos –acaso *todos* los procesos psicológicos– se descomponen en subrutinas efectivas que de alguna manera llevan a cabo nuestros cerebros. Ése es, en efecto, el núcleo de la argumentación ofrecida por Dennett (1978b), quien se sirve de la tesis de Church-Turing para alcanzar la conclusión de que toda teoría psicológica que no proceda a tal descomposición de los procesos que pretenda explicar, o a los que apele en el curso la explicación, en subrutinas efectivas, *ergo* Turing-computables, incurrirá *ipso facto* en una flagrante falacia homuncular –o bien en una ruptura con los principios mecanicistas. En palabras de Dennett:

> But now we can see that the supposition that there might be a non-question-begging non-mechanistic psychology gets you nothing, unless accompanied by the supposition that Church's Thesis is false. For a non-question-begging psychology will be a psychology that makes no ultimate appeals to unexplained intelligence, and that condition can be reformulated as the condition that whatever functional parts a psychology breaks its subjects into, the smallest, or most fundamental, or least sophisticated parts must not be supposed to perform tasks or follow procedures requiring intelligence. That condition in turn is surely strong enough to ensure that any procedure admissible as an "ultimate" procedure in a psychological theory falls well within the intuitive boundaries of the "computable" or "effective" as these terms are presumed to be used in Church's Thesis. The intuitively computable functions mentioned in Church's Thesis are those that "any fool can do," while the admissible atomic functions of a psychological theory are those that "presuppose *no* intelligence." If Church's Thesis is correct, then the constraints on mechanism are no more severe than the constraints against begging the question in psychology, for any psychology that stipulated atomic tasks that were "too difficult" to fall under Church's Thesis would be a theory with undischarged homunculi [...]. (Dennett 1978b: 83)

Si concedemos a Dennett la premisa implícita de que todos los procesos a los que nos es legítimamente dado apelar en una explicación psicológica "mecanicista" son procesos "mecánicos" en el sentido intuitivo acotado por Turing, el argumento es impecable. Resulta, sin embargo, que esa premisa es precisamente lo que se discute. Si no fuera cierto que todos los procesos psicológicos son Turing-computables –o, si se quiere, susceptibles de descomposición en procesos Turing-computables, aunque la diferencia es superflua–, la acusación de falacia por homuncularidad que blande Dennett descansaría sobre una falacia por equivocación: el atributo "mecánico" no tiene el mismo significado cuando, de acuerdo con el análisis de Turing, califica a un

procedimiento como algorítimico que cuando lo califica como admisible en una explicación científica a la que, por ceñirse a dicho requisito, denominamos "mecanicista". O, por reforzar la simetría de la réplica: la acusación de petición de principios se basa en una petición de principios. Desde una perspectiva epistemológica, la conclusión de razonamientos como el de Dennett –quien de hecho cita el trabajo pionero sobre la explicación psicológica de Fodor (1968)– no podía hacerse esperar: tal como planteaban Fodor (1981c: 130) o Boden (1988: 259), toda teoría psicológica habría de ser expresable en el escueto formalismo de Turing.

Con la vista puesta en el mismo presupuesto oculto del argumento de Dennett, Piccinini (2003, 2004, 2007) ha sostenido que, mientras que las explicaciones basadas en la apelación a procedimientos efectivos –mecánicos en el sentido de Turing– constituyen únicamente un subconjunto de las explicaciones mecanicistas, "[…] in the philosophy of psychology tradition that goes from Fodor to Dennett and beyond, explanation by appeal to effective procedures and mechanistic explanation have been conflated" (Piccinini 2007: 113). Si bien toda explicación mecanicista está sujeta a la exigencia de no reclutar, encubiertas en el *explanans*, las capacidades que conforman el *explanandum* –salvo, claro, que postule una estructura jerárquica que permita a su vez dar cuenta de esas capacidades mediante, como diría Dennett, homúnculos cada vez más estúpidos–, no es cierto que toda explicación mecanicista esté sujeta a la tesis de Church-Turing, puesto que los procesos o capacidades básicas a los que apele una explicación mecanicista bien pueden ser mecánicos en el sentido relevante sin ser procedimientos efectivos en el sentido de Turing –por ejemplo, porque sencillamente no sean computaciones. Lo que Piccinini trata de afianzar, en definitiva, es que la naturaleza computacional de los procesos psicológicos es un asunto empírico, que el trabajo de Turing o de Church no puede zanjar:

> [...] the empirical hypothesis that the brain is a computing mechanism [...] is not something that can be settled *a priori*. [...] To determine the relevance of procedures, whether effective or not, to neural or psychological theories, it seems more fruitful to develop and examine empirical theories of mind and brain rather than arguing *a priori* about these matters. (Piccinini 2007: 111)

En esto, el enfoque adoptado por Piccinini se aparta de lado a lado de la posición de Searle (1992), quien se quejaba precisamente de que la naturaleza computacional de los procesos psicológicos se pretendiera presentar como una cuestión *meramente empírica*, sobre la que no cabía análisis conceptual –un análisis que, a su entender, no arrojaba otra conclusión que la de que los procesos psicológicos *no* son procesos computacionales, pues aquellos son intrínsecamente semánticos y éstos, por definición, sintácticos, o formales (*cf. infra*)[123]. Coincide sin embargo Piccinini con

[123] Acaso algún indicio pueda, con todo, rastrearse en Piccinini de los planteamientos de Searle. El ejemplo de un proceso que es mecánico sin ser computacional aportado por Piccinini (2007: 115) –a saber, cualquiera de los que constituyen el funcionamiento del estómago– resulta ser el mismo en el que solía insistir Searle (1992: *passim*, *cf.* por ejemplo 1, 28-29, 90-91, 208, 224) a efectos de proporcionar para ciertas propiedades mentales una analogía desprovista de connotaciones computacionales: como

posiciones a la que, como a la de Pyslysyhn, se ha dado valor de ortodoxia cognitivista: que Pylyshyn también pretende tomar el asunto como cuestión empírica se infiere fácilmente de su argumento de que la única respuesta que conocemos a la pregunta de qué mantiene en pie el paralelismo entre "[…] the behavioral patterns caused by the physical instantiation of the representational states and the patterns captured by referring to the semantic content of these states" (Pylsyhyn 1984: 39) viene dada por la hipótesis de que el cerebro es un mecanismo computacional. Ahora bien: los términos en que Pylyshyn plantea la pregunta presuponen la ineficacia causal de la semántica, pero esa premisa –la de que "[…] the semantics of representations cannot literally cause a system to behave the way it does; only the material form of the representation is causally efficacious" (Pylyshyn 1984: 39)– no parece haber madurado como hipótesis empírica sino, al igual que en Searle, como resultado de un análisis conceptual –o acaso, cabría pensar en ambos casos, de posicionamientos preteóricos.

Vistas así las cosas, la universalidad de la máquina universal de Turing –su capacidad de remedar el funcionamiento de cualquier máquina de Turing– y la tesis de Church-Turing –que hay una máquina de Turing capaz de computar cualquier procedimiento que sea computable en el sentido intuitivo acotado por Turing– parecen al fin y al cabo haber dado lugar, recocidas en la misma cazuela, ni más ni menos que al mismísimo bálsamo de Fierabrás, a la panacea, si no de las dolencias de los mortales, sí al menos de sus problemas de cálculo o razonamiento: el longevo espejismo de la *characteristica universalis*, o *spécieuse générale*, que Leibniz (1666, 1678) –como apasionadamente ha relatado Yates (1966)– purgaría de los arcanos y sortilegios heredados por Raimundo Lulio de las artes antiguas de la memoria. En virtud de sus lazos con la teoría de autómatas y la teoría de la computación, el cognitivismo entroncaba –también lo ha sabido ver Rivière (1991b: 131)– con "[…] el ideal racionalista […] de un lenguaje automático y completo para el razonamiento"; de ese modo, además, se perfilaba como una restitución de la reflexión epistemológica que, salvo quizá durante el breve dominio conductista, ha sido inseparable del desarrollo histórico de la psicología –también en la tradición empirista que impulsa, como es bien sabido, los primeros pasos que de la mano de Wundt daría la psicología científica[124].

Una *máquina universal*: en el carácter universal de la razón humana había creído ver Descartes lo que la apartaba por igual de animales y de *otras* máquinas, pues no otra cosa eran los animales sino autómatas que, como estos, "[…] no obran

---

la digestión es al estómago, la consciencia sería al cerebro. Cabe también pensar, sin embargo, que tanto Piccinini como Searle no hacen sino prorrogar la vida de un símil al que ya había recurrido Cabanis (1802), y que Vogt (1847) y Ludwig (1852-1856) popularizarían (*cf*. *infra*).

[124] Es precisamente ese vínculo entre epistemología y psicología del que se sirve Quine (1969: 82-83, *supra*) para tratar de destilar la epistemología tradicional hasta que en la redoma sólo quede psicología –y después, claro, la propia psicología, hasta que sólo queden correlaciones entre estímulos y conductas descritos en un lenguaje fisicalista.

por conocimiento, sino por la disposición de sus órganos": "[…] mientras que la razón es un instrumento universal, capaz de servir en cualquier circunstancia, estos órganos necesitan una determinada disposición particular para cada acción" (Descartes 1637: 79). No es raro, entonces, que ante la perspectiva de rozar siquiera con las yemas de los dedos aquella *mathesis universalis* que –según la concebía Descartes, ya en 1628, en la primera de sus *Reglas para la dirección del espíritu*– permitiría al alma formarse "[…] juicios sólidos y verdaderos sobre todo lo que se le presenta", la incipiente concepción funcionalista de la mente, escudada por un aparato computacional que podía resultar tan arduo como fascinante, se viera espureamente avivada por un cierto enardecimiento. Lo que parecía a punto de revelarse ante nuestros atónitos ojos era ni más ni menos que, en palabras de Platón, "[…] las miríficas visiones que ofrece la intimidad de las sendas celestes, caminadas por el linaje de los felices dioses", e incluso, más allá de "[…] las empinadas cumbres, por lo más alto del arco que sostiene el cielo […] lo que está al otro lado" (*Fedro*: 247*a*-*c*)[125] –o, si se prefiere una metáfora más terrestre, el ancho mundo iluminado por el sol que ciega al prisionero que abandona la caverna (*República*: 514*a*-517*b*).

En efecto, los avances lógicos y matemáticos que se encarnaban en aquellas fascinantes máquinas de cómputo alentaban esperanzas la hondura de cuya raigambre racionalista es acaso atisbada por Pylyshyn (1984: 50) cuando apunta que dichos avances constituían, en cierto sentido "[… the] crystallization of the Platonic ideal of "pure form" divorced from all content". Idéntica geneología le atribuyen Dreyfus y Dreyfus (1988) –inspirándose en la inclemente crítica de la filosofía "tradicional" que Heidegger perfila en *El ser y el tiempo* (1927)– a la idea de construir toda nuestra teoría psicológica en torno al concepto de procesamiento de información. Ese proyecto no sería, a su entender, sino un hito más en la reiterada tendencia de cierta corriente de pensamiento filosófico –en la que son vecinos Platón, Descartes, Leibniz o Kant; incluso, con algún matiz, Aristóteles– a ignorar o distorsionar "[…] sistemáticamente el contexto cotidiano de la actividad humana" (Dreyfus y Dreyfus 1988: 355), equiparando comprensión, en cualquier ámbito, a posesión de una teoría, y posesión de una teoría a caracterización abstracta de relaciones entre elementos descontextualizados. Resulta seguramente innecesario enfatizar que la valoración que dicho proyecto merece para Dreyfus y Dreyfus es mucho más severa que la que le concede Pylyshyn (1984: 50, *supra*).

Más allá de lo que Pylyshyn intuye, o de lo que Dreyfus y Dreyfus, en la estela de Heidegger, tratan de impugnar, cabe ciertamente hallar signos de tal cristalización del concepto de estructura abstracta en las mismas páginas de *Crátilo* en las que hemos advertido *supra* los primeros vestigios de la constatación de que no es fácil construir una explicación del acto de nombrar que de cabida a la posibilidad de

[125] Un conocido grabado de Camille Flammarion para su propia *Météorologie Populaire* (1888: 163), en el que un hombre, a todas luces un astrónomo, asoma su cabeza por el exterior de la bóveda celeste, es quizá la representación visualmente más conspicua de la concepción platónica del conocimiento tal como ésta se refleja en *Fedro*. La descripción que acompaña al grabado es: "Un missionaire du moyen àge raconte qu'il avait trouvé le point où le ciel el la Terre se touchent…".

hacerlo erróneamente. En un esbozo temprano y tentativo de la teoría de las Ideas, Sócrates, discutiendo con Hermógenes en *Crátilo* 390*a*, deja, efectivamente, entrever la tesis de que un mismo orden abstracto puede encarnarse en distintas composiciones de elementos –alude a los nombres y las sílabas en distintos idiomas, pero también a los distintos fragmentos de hierro que usan distintos herreros para construir "el mismo instrumento"–, la misma tesis que hemos visto ir quedando bosquejada en las más tempranas reflexiones sobre lo que era y lo que no era mimético en los autómatas pensantes, que enseguida veremos florecer en la polémica entre Hilbert y Frege a cuenta de la naturaleza de las teorías axiomáticas, y que fertilizará la concepción funcionalista de lo mental a través del pensamiento de Turing y su exégesis en Putnam.

Las máquinas pensantes y la crisis de fundamentos de la matemática

Ni siquiera es fácil hoy en día –menos había de serlo entonces– calibrar hasta qué punto se encerraban cambios radicales, que afectarían a toda nuestra sociedad, en las investigaciones emprendidas por la generación de lógicos y matemáticos de la que formaban parte Turing y Church y en cuanto germinaría a partir de ellas, así que no es raro que entender cabalmente las repercursiones que podían tener sus trabajos sobre nuestra comprensión de nuestros propios procesos mentales y su relación con el funcionamiento de nuestros sistemas nerviosos siga siendo, si cabe, tarea aún más ardua, y que siga viéndose plagada de descuidos y confusiones.

Sea como sea, parte de las motivaciones iniciales de Turing, de índole metalógica, aparecen hoy muy distantes de los frutos que han dado: la principal aplicación teórica que Turing (1936, 1937) otorgaba a su computadora lógica mecánica no era sino la de construir una prueba de que *no existe* ningún procedimiento efectivo que determine, para cualquier expresión bien formada del cálculo de predicados, si dicha expresión es o no es un teorema en dicho sistema lógico –o, lo que dada la tesis de Turing viene a ser lo mismo, que *ninguna máquina de Turing puede hacer tal cosa*. Durante sus años de estudio en Cambridge, en las clases de Max H.A. Newman sobre fundamentos de las matemáticas, Turing se había familiarizado con la lista de problemas matemáticos irresueltos que David Hilbert había presentado en 1900 al Congreso Internacional de Matemáticos, en La Sorbona, así como con las tres cuestiones básicas que había planteado en 1928, en Bolonia: si las matemáticas son completas, si son consistentes, y si son decidibles.

El programa formalista de fundamentación de las matemáticas que subyacía a las preguntas de Hilbert se había visto ya arrasado en 1931 –el año en que la solicitud de admisión del joven Turing era rechazada en Trinity Collage pero aceptada en King's Collage– por la publicación del teorema de incompletitud de Kurt Gödel. Con sólo veinticinco años, apenas unos meses después de haber mostrado en su tesis doctoral que el cálculo de predicados es un sistema completo en el sentido de Hilbert, Gödel conmocionó a la comunidad matemática al demostrar que cualquier

formulación de las matemáticas que podamos construir siguiendo un procedimiento efectivo, que tenga al menos el poder expresivo de la aritmética de Peano, y que resulte consistente, será forzosamente incompleta, pues nos permitirá construir enunciados verdaderos que no se puedan demostrar, ni refutar, mediante los recursos del propio sistema –sus *enunciados de Gödel*. En otras palabras, el joven Gödel había establecido que las dos primeras preguntas de Hilbert no podían recibir una respuesta afirmativa para el mismo sistema matemático, supuesto que éste pueda expresar la aritmética elemental y haya sido generado de forma efectiva –es decir, que el conjunto de sus axiomas sea enumerable recursivamente. La tercera de las preguntas formuladas por Hilbert en Bolonia –si las matemáticas son decidibles– se conocía ya como el problema de la decisión, o *Entscheidungsproblem*: según lo describían Hilbert y Ackermann (1928/1938: 21-22), dicho problema quedaría resuelto si contáramos con un procedimiento efectivo para decidir, para cualquier expresión lógica dada, si ésta es verdadera o falsa[126]; a su juicio "[…w]e are justified in calling it the main problem of mathematical logic" (Hilbert y Ackermann (1928/1938: 113).

Acotado para el caso de la lógica de primer orden –que el problema tiene una solución afirmativa en el marco del cálculo de proposiciones era ya conocido–, éste era el blanco del ataque de Turing, como lo había sido del de Church: ¿existe un algoritmo capaz de decidir si un enunciado cualquiera del cálculo de predicados es un teorema? La "observación sobre el *Entscheidungsproblem*" mencionada en el título del artículo de Turing de 1936 era, ni más ni menos, que la irresolubilidad para el cálculo de predicados del problema avanzado por Hilbert se deriva, como corolario, de la demostración de que no puede existir una computadora lógica mecánica –es decir, un procedimiento efectivo– capaz de predecir si otra computadora lógica mecánica dada –conocidos, pues, el contenido de la tabla de máquina y de la cinta– se detendrá o no. Que la solución al problema de la parada es negativa es –*in nuce*– fácil de comprender: supongamos que existiera una máquina *E* que arrojara el resultado *1* siempre que procesa los datos relativos a una máquina que sí se detiene, y el resultado *0* cuando se enfrentase a una máquina que no se detiene. Podríamos entonces programar otra máquina *D*, que reprodujera el algoritmo de *E* para decidir si una máquina dada se detiene o no, y entonces hiciera lo siguiente: detenerse si el resultado es *0* –es decir, si la máquina analizada no se detiene– y entrar en un bucle sin fin si el resultado es *1* –o sea, si la máquina analizada se detiene. Pero si entonces le proporcionáramos a la máquina *D* su propio código, ésta debería detenerse siempre que no se detuviera, y no detenerse siempre que se detuviera[127]. La

[126] O más exactamente, su *validez universal* (si es verdadera en cualquier sistema apropiado) y su *satisfactibilidad* (si lo es en alguno).

[127] En el abordaje del problema de la parada ensayado por Turing, al igual que en los enunciados de Gödel, en la paradoja de Russell sobre la clase formada por aquellas clases que no pertenecen a sí mismas (*cf. infra*), y en el argumento diagonal de Georg Cantor (1891) –que sirvió de inspiración a Russell, Gödel y Turing–, aflora vigoroso el recuerdo de la célebre paradoja del mentiroso, que la tradición, desde Diógenes Laercio (*Vidas* II: 108), atribuye a Eubúlides de Mileto, discípulo aventajado de Euclides de Megara: un hombre que afirma que miente, ¿miente? –si miente, dice la verdad; si dice

contradicción nos alerta –*reductio*– de que el supuesto inicial era insostenible: no puede existir una máquina *E* que decida para cualquier otra máquina si ésta se detendrá o no. La respuesta al problema de la parada es: no. Ahora bien, si ninguna máquina de Turing puede decidir si un procedimiento efectivo dado –la tabla de máquina de otra máquina de Turing– llegará a detenerse cuando se enfrente a determinados datos –el contenido de la cinta de esa máquina–, tampoco es difícil ver, al menos intuitivamente, que es imposible que exista "[…] un procedimiento efectivo para decidir, para cualquier expresión lógica dada, si ésta es verdadera o falsa" Hilbert y Ackermann (1928/1938: 21-22, *supra*). La razón es que si existiera tal procedimiento efectivo, podríamos implementarlo en una máquina de Turing, pero esa máquina sería precisamente la máquina *E*, cuya imposibilidad hemos comprobado. El problema de la decisión se ha convertido, dada la tesis de Turing, en el problema de la parada. *Ergo* la respuesta al problema de la decisión (en el marco del cálculo de predicados) es: no.

En suma, la utilidad más temprana de las máquinas ideadas por Turing fue precisamente demostrar que la decisión de si una expresión lógica cualquiera del cálculo de predicados es verdadera o falsa no es computable –no es Turing-computable ni, por tanto, según la tesis de Turing, computable en sentido intuitivo–, como tampoco lo son muchos otros procedimientos y funciones que, no obstante, bien podemos considerar ordenados, sistemáticos, etc. No deja de resultar irónico que *ese* logro de Turing haya acabado dando pie a la especie de que todo proceso al que pueda darse una descripción científica –física, mecanicista, lógica, matemática, etc.– sería Turing-computable.

De hecho, el programa formalista de Hilbert –el sueño de Hilbert, si se quiere– constituía a todas luces la expresión última de las ambiciones racionalistas de Descartes o Leibniz: la lengua perfecta (*cf*. Eco 1993), el cálculo sublime, oracular, que habría de permitir conocer todas las respuestas con sólo formular debidamente las preguntas (*supra*), y cuyo germen también Bronckart (2002: 388) reconoce en Platón[128]. Es quizá un signo del embrujo que ese proyecto ha ejercido sobre nuestra tradición intelectual que hayamos llegado a confundir la constatación de que la

---

la verdad, miente. Si Laercio (*Vidas* VII: 196-198) está en lo cierto, Crisipo habría dedicado no menos de catorce libros a la elucidación de la paradoja, todos ellos –como la casi totalidad de la obra del prolífico alumno de Zenón, perdidos hoy. Hasta principios del s. XX, entonces, nos habría eludido la comprensión del carácter medular de dicha paradoja –que Séneca (*Cartas morales a Lucilio* XLV: 101) consideraba una pérdida de tiempo– de cara a la fundamentación de la lógica y la matemática. De hecho, Russell (1908) introduce la paradoja que lleva su nombre, y que había descubierto siete años antes, al hilo de una de las primeras referencias modernas a la paradoja de Epiménides el cretense, a quien se atribuye la afirmación (cuya intención era sin duda moralizante, y en absoluto paradójica) de que los cretenses mienten siempre –una variante estilística de la paradoja del mentiroso que, no obstante, arrastra algunas dificultades lógicas a las que la versión de Eubúlides es inmune (*cf.* Rescher 2001, Sorensen 2003).

Una obra de divulgación en la que se da a un amenísimo tratamiento a varias paradojas y acertijos lógicos de índole autorreferencial es Smullyan (1978).

[128] Precisamente en el *Crátilo*, al que se ha dado *supra* una lectura un tanto diferente.

lengua perfecta es una quimera con la promesa de su advenimiento. Dicha promesa, además, habría llegado –si lo hubiera hecho– en momentos de zozobra: parte de lo que impulsó a Hilbert a retar a la comunidad matemática con sus tonificantes veintitrés problemas irresueltos, y a bosquejar el programa formalista de una matemática completa, consistente y decidible –que, aunque fallido, daría su forma moderna a la reflexión metamatemática– era su profunda indignación ante las posiciones que había propagado du Bois-Reymond (1872 *apud* Bueno 1990: 69, *supra*), las cuales, aunque expresadas con pretendida "viril renuencia" en su *¡Ignorabimus!*, no eran para Hilbert otra cosa que un imperdonable signo de abatimiento. Tan clara es la huella de las conclusiones de du Bois-Reymond en el ánimo de Hilbert que el desafío que lanzara a sus colegas en París –como concienzudamente anota Tennant (2007)– toma casi como lema su negación:

> This conviction that each and every mathematical problem is soluble is a powerful incentive to us in our work. We hear within us the constant call: *There is the problem. Seek its solution. You can find it by pure reason, for in mathematics there is no 'Ignorabimus'.* (Hilbert 1901: 298 *apud* Tennant 2007: 18)

No parece descabellado, de hecho, conjeturar que la resignación que du Bois-Reymond venía aconsejando rozaba cada vez más una llaga abierta para Hilbert. Sin duda había motivos para albergar ciertas vacilaciones en torno a la posibilidad de dar una fundamentación firme, en el sentido anhelado por Hilbert, a la totalidad del conocimiento matemático, vacilaciones cuyo repudio exacerbaría, en el ánimo de Hilbert, la actitud de du Boys-Reymond. Debe consignarse, por una parte, el temprano conocimiento que Hilbert había tenido de la paradoja de Russell y de sus devastadores efectos sobre la teoría de conjuntos de Frege –Ernst Zermelo parece haber descubierto la paradoja poco antes que Russell, y habérsela comunicado a Hilbert en Göttingen (*cf.* van Heijenoort 1967: 124), más o menos por las mismas fechas en las que Russell remitiera su famosa carta a Frege del 16 de junio de 1902. Además, desde luego, está la noticia del descubrimiento de las geometrías consistentes no euclídeas[129], que había hecho trepidar las hipótesis que subyacen a la propia geometría –la expresión es la que el propio Riemann usara para titular la tesis de habilitación que bajo la supervisión de Carl F. Gauss había leído en 1854, también en Göttingen: *Über die Hypothesen, welche der Geometrie zu Grunde liegen*[130]. No en vano, el cuarto de los veintitrés problemas inventariados por Hilbert, por ejemplo, tenía que ver con la necesidad de generalizar la noción de línea recta al contexto de espacios curvos, analizando exhaustivamente las propiedades de las geodésicas bajo diferentes métricas. La llaga abierta en el ánimo de Hilbert era, de hecho, la primera

[129] Es decir, en las que no se verifica, por ejemplo, el quinto postulado de los *Elementos* de Euclides, a saber, en la moderna formulación de Playfair (1860): que para toda línea *L* sobre un plano y todo punto *P* por el que no pase *L* hay una única línea que pasa por *P* sin cruzar *L*, a saber, la paralela de *L*. La negación del axioma de las paralelas es la base de las geometrías de Lobachevsky (1829) y Bolyai (1832).

[130] Y que sólo tras la muerte de Riemann publicaría Richard Dedekind (Riemann 1867).

crisis de fundamentos –*Grundlagenkrise*– de la matemática moderna, tan devastadora y tan fructífera como la que propició el descubrimiento de magnitudes inconmensurables y números irracionales cuya revelación fuera de los círculos secretos pitágoricos atribuye la tradición a Hípaso de Metaponto.

El esfuerzo por adoptar el método axiomático a las nuevas geometrías no euclídeas fue, de hecho, una de las principales preocupaciones de Hilbert, y lo que le llevó a su acre controversia con Gottlob Frege: en efecto, la idea de que "[…] los axiomas son unos meros esquemas abstractos, que en sí mismos no son verdaderos ni falsos" resultaba inaceptable –como con cierta mordacidad recuerda Mosterín (1984: 177)– "[…] para los defensores de la concepción clásica del método axiomático, no sólo para los espíritus obtusos (como ya había previsto Gauss), sino incluso para mentes tan agudas como la de Frege"[131]. Acerca de la naturaleza de las teorías construidas en el seno del método axiomático, la propuesta radical de Hilbert –como le escribía a Frege en una carta fechada el 29 de diciembre de 1899– es que:

> Cada teoría no es sino un tinglado o esquema de conceptos junto con ciertas relaciones necesarias entre ellos, y sus elementos básicos pueden ser pensados arbitrariamente. Si entiendo por puntos, etc., cualquier sistema de cosas, por ejemplo, el sistema formado por amor, ley, deshollinador, etc., y considero que todos mis axiomas resultan válidos para esas cosas, entonces también resultan válidos para esas cosas mis teoremas, como, por ejemplo, el de Pitágoras. Con otras palabras: cada teoría puede ser aplicada a una infinidad de sistemas de elementos básicos. (Hilbert 1899 *apud* Frege, *Wissenschaftlicher Briefwechsel*: 67 *apud* Mosterín 1984: 180)

No hay que extrañarse de que las palabras de Hilbert evoquen vivamente, cuando no las provocativas proclamas funcionalistas acerca de las múltiples encarnaciones que podría en principio adoptar lo mental –como cuando Putnam aseguraba que, por cuanto atañe a nuestra psicología, "[...w]e could be made of Swiss cheese and it wouldn't matter" (Putnam 1975b: 291, *supra*)–, sí al menos las expresiones más comedidas de la misma idea. Así, por ejemplo, el propio Putnam, con ánimo más temperado, reiteraría que:

> The functional organization (problem solving, thinking) of the human being or machine can be described in terms of the sequences of mental or logical states respectively (and the accompanying verbalizations), without reference to the nature of the "physical realization"of these states. (Putnam 1975: 373)

[131] En su lúcido estudio, Mosterín (1984) no duda en enlazar la confianza inquebrantable de Frege en la geometría euclídea con "[…] su concepción kantiana de la geometría, cuyos axiomas se captarían por una intuición pura del espacio" (Mosterín 1984: 179). El propio Mosterín apunta también, en un trabajo sobre "Kant como filósofo de la ciencia" que los juicios sintéticos *a priori* son posibles en la geometría euclídea, de acuerdo con la *Crítica de la Razón Pura*, porque ésta "[…] es la teoría del espacio euclídeo, que es la forma que nuestra sensibilidad impone a todo objeto al percibirlo"; desde luego, "[…n]o es que el mundo real sea euclídeo […] (eso sería una mera afirmación metafísica […])", "[…] lo que es euclídeo es el mundo perceptual, apariencial" (Mosterín 1984: 153). Sin embargo, dado que Kant –como es bien sabido– pretende fundar el conocimiento matemático sobre las intuiciones que rigen la percepción, se seguiría de su análisis la imposibilidad de toda geometría no euclídea.

En un tono parecido, aunque ceñido provisionalmente al ámbito de los sucesivos estados de cualquier computadora finita –es decir, desprovisto de alusiones a lo mental–, Pylyshyn, que apuntaría también que "[…n]o existen límites a las combinaciones de propiedades físicas que se puedan usar para instanciar, digamos, el *modus ponens*" (Pylyshyn 1984: 140), señala que:

> Not only can such computational sequences be realized in any digital computer ever made or that will ever be made, but they can be realized in devices operating in any imaginable media –mechanical (as in Charles Babbage's Analytical Engine), hydraulic, acoustical, optical, or organic –even a group of pigeons trained to peck as a Turing machine! (Pylyshyn 1984: 55-57)

Tal vez pensando en las máquinas de Babbage, o en los múltiples modelos hidráulicos y mecánicos de procesos psicológicos que –como venimos viendo– antecedieron a las primeras computadoras, y que él conocía cuando menos a través de su lectura de Craik (1943), Johnson-Laird (1988) acentuaría la misma conclusión que Pylyshyn: la máquina en la que se encarne un programa determinado

> […] podría hacerse a base de engranajes y manivelas, como las viejas máquinas de calcular mecánicas; podría construirse a base de un sistema hidráulico, a través del cual corriese el agua. Podría hacerse con transistores impresos en un chip de silicona, por el cual fluyese la corriente eléctrica. Incluso podría llevarse a cabo en el cerebro. Cada una de estas máquinas utiliza un medio diferente para representar los símbolos binarios, bien sea la posición de los engranajes, la presencia o ausencia de agua, el nivel del voltaje o, quizá, los impulsos nerviosos. (Johnson-Laird 1988: 43)

Vienen también al caso las palabras que Fodor (1985: 15, *supra*) dedica a glosar cómo la idea de clasificar instancias de estados psicológicos en tipos únicamente en virtud de sus relaciones funcionales nos permitió acogernos al ímpetu de:

> […] the emerging intuition that the natural domain for psychological theory might be physically heterogeneous, including a motley of people, animals, Martians (always, in the philosophical literature, assumed to be silicon based), and computing machines.

Personas, animales, alienígenas o autómatas, entonces, quedarían descritos por una única teoría psicológica, de la que todos esos diversos sistemas constituirían modelos: la teoría no es sino un "[…] esquema de conceptos junto con ciertas relaciones necesarias entre ellos […]" (Mosterín 1984: 180, *supra*); así pues, cualquier sistema de cosas que satisfaga las relaciones entre elementos establecidas sería también un modelo de la teoría.

Es precisamente el temor a a la conclusión hilbertiana de que "[…] los elementos básicos pueden ser pensados arbitrariamente" (Mosterín 1984: 180, *supra*) lo que –como luego veremos con más detenimiento– llevará a Block,como veremos, a enrocarse en la necesidad de que tanto estímulos como respuestas queden descritos

en términos físicos y no, como los estados internos del sistema, en términos funcionales, ya que –piensa Block:

> […] characterizing inputs and outputs themselves functionally would appear to yield an abstract structure that might be satisfied by, for instance, the economy of Bolivia under manipulation by a wealthy eccentric [...] (Block 1996: 24)

Los reparos de Block, de hecho, no pueden sino recordar a los que Frege planteaba en su reseña de los *Grundlagen der Geometrie* de Hilbert, donde, tras asemejar las teorías abstractas bosquejadas por Hilbert a sistemas de ecuaciones con varias incógnitas –en las que cada solución del sistema de ecuaciones es un "sistema de cosas" que conforma un modelo de la teoría–, Frege protesta retóricamente: "¿Quién nos dice que este sistema de ecuaciones tenga alguna solución y que ésta sea unívoca?" (Frege 1903: 370 *apud* Mosterín 1984: 182).

Incluso el aire deliberadamente disparatado de la enumeración de los elementos de un tentativo modelo de la geometría euclídea con la que Hilbert sin duda erizaría el hosco temperamento de Frege –"amor, ley, deshollinador, etc."– encuentra su remedo en algunas formulaciones contemporáneas de la idea de que una misma psicología puede darse en diversas encarnaciones. A pesar del tono satírico –o precisamente en virtud de él–, la semejanza se distingue con especial claridad en un párrafo en el que Searle atestigua que (según quienes emplean la tesis de realizabilidad múltiple como sustento de cierta concepción de la inteligencia artificial, contra la que él va a presentar sus categóricas objeciones), la máquina en la que llegado el caso hiciéramos funcionar un programa que simulara un determinado estado psicológico (la sed, en el ejemplo elegido por Searle), y que por lo tanto (siempre según esa posición, que Searle bautiza como "Inteligencia Artificial Fuerte") literalmente sentiría sed, bien podría ser, en lugar de una computadora más o menos convencional,

> [...] an ant colony (one of their examples), a collection of beer cans, streams of toilet paper with small stones placed on the squares, men sitting on high stools with green eye shades –anything you like.
>
> So let us imagine our thirst-simulating program running on a computer made entirely of old beer cans, millions (or billions) of old beer cans that are rigged up to levers and powered by windmill. (Searle 1982: 462)

El paralelismo era, a fin de cuentas, esperable: en efecto, como ya hemos visto, es en el trabajo de Turing donde Putnam halla los recursos conceptuales para resguardar del riesgo de circularidad la lección acerca del carácter relacional de lo mental que habíamos aprendido del conductismo lógico y, de esa forma, es en el trabajo de Turing donde germina la noción de que sistemas materialmente dispares pueden ser funcionalmente –*ergo* psicológicamente– equivalentes. Así, fue el propio Turing quien adelantándose a las reflexiones de Putnam dejara apuntado que:

> A menudo se da importancia al hecho de que las computadoras digitales modernas son eléctricas y que el sistema nervioso también lo es. Puesto que la máquina de Babbage no era eléctrica y puesto que todas las computadoras digitales son equivalentes en cierto sentido, observamos que este uso de la electricidad no puede tener importancia teórica. […] Entonces, el hecho de que se utilice electricidad resulta tan sólo una semejanza muy superficial. Si realmente deseamos encontrar tales semejanzas, deberíamos buscar analogías matemáticas en el funcionamiento (Turing 1950a: 59)

Las investigaciones de Turing, por lo demás, están tan hondamente entrelazadas con Hilbert que nada hay de digno de asombro en el hecho de que las nociones hilbertianas de teoría abstracta y modelo de una teoría anticipen con toda claridad la idea de las relaciones entre la teoría psicológica y sus diversos modelos –encarnaciones, realizaciones– que amartillará los fundamentos del funcionalismo. Una interpretación parecida puede encontrarse en Tennant (2007: 24-25):

> It was Hilbert who first developed our understanding of a set of axioms as implicitly defining the theoretical primitives involved[132]; and who stressed that as far as matters of interpretation were concerned, all that counted was structure, rather than the actual individual objects involved. A geometrical theory, for example, could just as well be interpreted as being about tables, chairs and beer mugs, as about points, lines and planes. Thus the theoretical "objects" involved could be abstracted out as mere "loci of relational intersections". This mind-set no doubt contributed decisively to philosophers' later construal of mental states as logical states of a program, which could be identified by their "place within a transitional network", rather than by the physical condition of whatever biological organ or substrate produced the mental functioning in question.

En el análisis de Tennant, el desarrollo de una noción plenamente abstracta de las entidades teóricas por parte de Hilbert corre parejo al de un concepto igualmente abstracto de función –para Lejeune Dirichlet (1837), un conjunto ordenado de pares tal que dos de ellos cualesquiera coinciden en su segundo elemento siempre que coinciden en el primero[133]– para hacer posible que comencemos a pensar en términos abstractos también sobre lo mental, entendiéndolo como *función* del sistema nervioso[134]. Entender la función psicológica como algo desligado de su substrato neurológico sería posible, entonces, porque:

---

[132] Si bien parece que el propio Hilbert no llegó a utilizar la noción de definición implícita: *cf.* Mosterín (1984: 182).

[133] Aunque tal formulación explícita no se debe al propio Dirichlet, sino a Hankel, quien de hecho la bautizó como "el concepto de función de Dirichlet" para distinguirlo del concepto anterior, ligado a una expresión matemática o lógica –i.e., lingüística– determinada: *cf.* Calinger (1996: 318).

[134] Aun así, como ha quedado visto *supra*, la concepción funcionalista de lo mental que encontramos en Fodor o en Pylyshyn no es hilbertiana hasta las últimas consecuencias: no se presenta como una teoría puramente formal, desligada de la realidad en el sentido de Schlick (1918/1925) –es decir, carente de todo compromiso ontológico más allá de la austera afirmación de que si existieran entidades que verificasen los axiomas de la teoría, ésta constituiría una descripción adecuada de dichas entidades–, sino como una teoría que incluye un juicio existencial rotundo –existen los estados mentales–, acompañado de una tesis sobre su clasificación funcional. En otras palabras: la de Fodor o Pylyshyn es una concepción funcionalista *y realista* de lo mental.

> [...t]o the extent that mathematical functions could be 'alinguistic', that is, not tied down to expression by any linguistic rule, so too could higher organic or cerebral functions be 'abiological', that is, not tied down to expression (or realization) by any particular anatomical structures. (Tennant 2007: 24)

Ésa sería –piensa Tennant– la concepción de lo mental que acabaría madurando en el funcionalismo cognitivista, y que haría mostrado sus primeros brotes en el pensamiento de Ernst Haeckel, cuando el ferviente darwinista de Potsdam, desbordando el craso concepto de superveniencia que habrían trazado Tyndall y Huxley –y en una obra titulada, en clara referencia a ellos y a Du Bois-Reymond, *Enigmas del Universo*–, apunta que:

> […f]urther developed brain operations, the formation of chains of reasoning that hang together, abstraction and concept-formation, the expansion of cognitive understanding by means of the plastic activity of imagination, and finally consciousness, thinking and philosophizing, are just as much functions of the ganglion cells or neurons of the cerebral cortex as are the [...] simpler mental activities. (Haeckel 1899: 13-14 *apud* Tennant 2007: 26)

Con independencia de que esté o no justificado leer en los textos de Haeckel precoces convicciones funcionalistas acerca de la relación entre lo mental y lo físico –los argumentos de Tennant no parecen del todo terminantes–, sí resulta razonablemente claro que tales convicciones guardan cierta continuidad con el esfuerzo de abstracción del que Hilbert aparece como adalid. La propia analogía entre los procesos psicológicos y los procesos computacionales, que a menudo se ha tomado como divisa del cognitivismo –la metáfora del ordenador, o más bien su interpretación literal, que ha impulsado entre otros Pylyshyn (1984: *xiii*, *infra*)– encuentra sus cimientos en las nociones de teoría abstracta y modelo vislumbradas por Hilbert. La tesis de Pylyshyn, de acuerdo con la cual la cognición *es*, en sentido plenamente literal, un tipo de computación, bien podría considerarse un caso particular de lo que Hilbert establece con carácter general acerca de la relación entre teorías y modelos. Como nos recuerda Mosterín (1984: 180) en relación con el foco original de la controversia –la axiomatización de la geometría euclídea y de las nuevas geometrías no euclídeas– en la propuesta de Hilbert, a diferencia de lo que ocurría en la kantiana:

> […] no se hace uso de intuición ni conocimiento previo ninguno, sino que lo único que es lícito suponer de los puntos, rectas y planos, etc., es lo que explícitamente se dice de ellos en los axiomas. A la inversa, cualquier sistema de cosas de las que se pueda decir lo mismo que los axiomas dicen de los puntos, rectas, etc., puede considerarse como un modelo de la geometría.

En efecto, *mutatis mutandi*: cualquier sistema de cosas de las que se puede decir lo mismo que los axiomas (de una hipotética teoría psicológica madura: axiomatizada) dicen de las representaciones internas, reglas de transición, etc., puede considerarse

como un modelo de la psicología. Los procesos cognitivos, pues, son un caso particular de procesos computacionales: el substrato hilbertiano del pensamiento de Pylyshyn parece rotundo. El propio Pylyshyn, de hecho, apunta algo lacónicamente que:

> It was the development of the notion of the universality of formal mechanism, first introduced in the work on foundations of mathematics in the 1930s, that provided the initial impetus for viewing the mind as a symbol-processing system. (Pylyshyn 1984: 51)

Antes incluso de que al conductismo comenzara a faltarle el aire, esa idea de abstracción funcional había venido enraizándose –como hemos visto con cierto detenimiento– al hilo de los cada vez más reiterados ensayos de replicar el comportamiento animal o humano en modelos mecánicos. Alguna señal hemos encontrado también –por ejemplo en Deutsch (1954: 7, *supra*)– de que en el trasfondo de esa incipiente lectura funcionalista de los avances en simulación acaso estuvieran involucradas reverbereraciones más o menos nítidas del pensamiento de Hilbert. Es innegable, con todo, que parte del ímpetu que había de adquirir el funcionalismo en los años venideros pudo quizá provenir sencillamente de la naturalidad con la que un variopinto bestiario de autómatas fue anidando en laboratorios, publicaciones, congresos, e incluso ferias y exposiciones, y de las reflexiones que la familiaridad con toda aquella fauna artificial pudo suscitar, sin que mediara en muchos casos noción alguna de los nexos que estas reflexiones pudieran mantener con una concepción formalista de la matemática.

Como nos hace ver Mosterín (1984: 182-183), por otro lado, hay un aspecto técnico de la controversia entre Hilbert y Frege a cuenta del método axiomático en el que el desarrollo de la lógica y la matemática moderna ha dado la razón a este último. Dado que Hilbert usa indistintamente el término "definición" para referirse a definiciones explícitas, estipulativas, como a la delimitación de los conceptos de *punto*, *recta*, *plano*, etc. por medio de los axiomas, Frege le reprocha la confusión en la carta que le remite el día de Reyes de 1900: "Me parece que lo que usted en realidad quiere definir son conceptos de segundo orden, pero que usted no los distingue claramente de los de primer orden" (Frege, *Wissenschaftlicher Briefwechsel*: 67 *apud* Mosterín 1984: 182-183). Exactamente: tal como señala Mosterín (1984: 183), los axiomas hilbertianos no definen los conceptos de *punto*, *recta*, *plano*, etc. –y menos en el mismo sentido en que uno puede definir explícitamente el concepto de triángulo–, sino que definen la *estructura abstracta del espacio euclídeo*, y lo hacen implícitamente. Pues bien: también aquí la concepción funcionalista de lo mental se presenta con el ropaje de una teoría abstracta en el sentido de Hilbert –con las salvedades, si se desea, apuntadas por Block (1996). Como se ha repasado al comentar el esbozo de especifición funcional del dolor de Putnam (1967a: 229, *supra*), el funcionalismo tiende a considerar los estados mentales –o los estados de tabla de máquina de un autómata, a los que Putnam entonces los asimila– como "estados de segundo orden", definidos por su posesión de propiedades físicas (de primer orden) en determinadas

relaciones recíprocas; los estados físicos definidos por la posesión de dichas propiedades de primer orden se consideran implementaciones de aquellos estados funcionales (de segundo orden). Más sucintamente, al hilo de la discutida cuestión de la eficacia causal de lo mental, se ha entrevisto ya como Braddon-Mitchell y Jackson (1996: 264, *supra*) se refieren a las propiedades psicológicas como "[…] second-order properties of having a property that plays a certain causal role". Que entender los conceptos psicológicos como conceptos de segundo orden sea explicativamente adecuado, que permita dotar a los estados psicológicos de la eficacia causal de la que el funcionalismo ha hecho gala de dotarlos, que nos autorice a postular *propiedades* o *estados* de segundo orden y no sólo conceptos de tal naturaleza, etc., son asuntos extremadamente espinosos, sobre los que habrá ocasión de regresar. Por el momento, es de esperar que tomar nota del modo en que la concepción funcionalista de lo mental arraiga en un terreno aparentemente tan apartado como el de la polémica sobre el método axiomático facilite el discernimiento de las aristas de esas y otras controversias contemporáneas. Desde luego, tal constatación no resta ningún valor a la intuición crucial que madura en la lectura de Turing por parte de Putnam: que es posible desligar nuestra idea de actividad psicológica de nuestra idea de actividad cerebral sin vernos por ello forzados a ceder a una concepción dualista de lo mental y lo físico.

Mentes y máquinas: metáforas de una metáfora

Una consecuencia inmediata de la aplicación de los principios de la teoría de autómatas a la tarea de explicar la naturaleza de los estados mentales –en realidad, de su mera identificación con estados caracterizados funcionalmente, aunque se prescinda de las herramientas importadas por Putnam–, es que éstos quedan delimitados de una manera que, más allá del mero compromiso con su existencia, resulta ontológicamente inocua. Esto –venimos de argumentar– enlaza al funcionalismo con la noción de teoría abstracta desarrollada por Hilbert en el contexto de la reformulación del método axiomático inspirada por el descubrimiento de las geometrías no euclídeas. Así, dado que la definición funcional de un estado mental no hace referencia alguna a sus propiedades físicas, sino únicamente a sus relaciones con aferencias, eferencias y otros estados mentales, cualquier sistema, sea cual sea su composición o estructura, en el que encontremos estados que se plieguen a ese patrón de relaciones habrá de ser incluido en el distinguido círculo de los sujetos psicológicos –los "objetos con mente", diría Rivière (1991)–, o excluido en virtud de algún principio que queda pendiente de arbitrar sin levantar sospechas de estarlo haciendo *ad hoc*.

Es éste el razonamiento que anima la punzante advertencia de que el funcionalismo, y por su mediación la psicología cognitiva, no se vale –como suele decirse– de una *metáfora computacional* para tratar de entender la mente, como antes se emplearon tantas otras, desde la tablilla de cera hasta la piedra luminiscente de

Bolonia –*cf*. Draaisma (1993). Al contrario, tanto la psicología cognitiva como la concepción funcionalista de la mente sobre la que esta disciplina se cimenta estarían comprometidos *velis nolis* con la tesis *literal* de que la cognición *es* un proceso computacional, o, con espíritu más polémico, de que una computadora adecuadamente programada *es* una mente. Particularmente tajante al respecto, como hemos visto, se muestra Pylyshyn (1984):

> […M]y proposal amounts to a claim that cognition *is* a type of computation. (Pylyshyn 1984: *xiii*)

Es decir: Pylyshyn interpreta la relación entre computación y cognición no como la de vehículo a tenor (Richards 1936), o sujeto secundario a sujeto primario (Black 1962) de una metáfora, sino más bien como la de clase a instancia, o tipo a caso. Lo que se propone es, entonces, una lectura literal de la homología entre ambas nociones, lectura cuya primacía el propio Pylyshyn atribuye a Newell y Simon (1972): en efecto, en el epílogo de *Human Problem Solving* queda dicho con toda rotundidad que "[…] programmed computer and human problema solver are both species belonging to the genus I[nformation] P[rocessing] S[ystem]". El razonamiento sobre el que reposa esta interpretación orbita en torno a la idea de representación interna, y más en particular a la de procesamiento de representaciones internas encarnadas en forma de símbolos. La tesis de que la actividad cognitiva consiste en procesamiento de información equivale –insiste Pylyshyn– a la de que la cognición no es, en sentido plenamente literal, sino computación:

> One of the central proposals that I examine is the thesis that what makes it possible for humans (and other members of the natural kind *informavore* [Miller 1984]) to act on the basis of representations is that they instantiate such representations physically as cognitive codes and that their behavior is a causal consequence of operations carried out on these codes. […T]his is precisely what computers do […] (Pylyshyn 1984: *xiii*)

Idéntico énfasis en la idea de que la relación entre cognición y computación postulada por el cognitivismo no es de orden figurado puede encontrarse en Cummins (1989). De nuevo, el vínculo entre una y otra radica en la noción de representación, y del vigor de ese vínculo se hace depender la capacidad explicativa de lo que, no en vano, se denomina teoría computacional de la cognición:

> It is an absolutely central thesis of the C[omputational] T[heory of] C[ognition] that representation in cognitive systems is exactly the same thing as representation in computational systems generally. This is why computation and representation can be thought of as independently available to *explain* cognition. (Cummins 1989: 117-118)

La misma tesis, desde luego, puede formularse con un aire más provocador: sólo es preciso hacer hincapié en que la clase de los "sistemas computacionales" abarcaría, junto a nosotros mismos, a máquinas a las que nos resistimos a considerar semejantes a nosotros siquiera en algún aspecto relevante –el que defina la pertenencia a dicha

clase–, como si de esa similitud se derivase *ipso facto* que nada pueda diferenciarnos de ellas. Pero es ahí –insiste Cummins– donde reside la fuerza explicativa que la noción de computación pueda conferir al cognitivismo:

> The C[omputational] T[heory of] C[ognition] proposes to explain cognitive capacities by appeal to representation and computation in exactly the way that the arithmetic capacities of calculators (such as addition) are standardly explained by appeal to representation and computation. The main explanatory strength of the C[omputational] T[heory of] C[ognition] is that it proposes to explain cognition in terms of antecedently understood notions of representation and computation –notions the C[omputational] T[heory of] C[ognition] takes to be unproblematic in a variety of familiar noncognitive contexts such as calculating and elementary computer programming. (Cummins 1989: 89)

En términos muy parecidos se expresa Horst (1996: 62) sobre la relación entre procesos computacionales y procesos psicológicos que postula la "teoría computacional de la mente", cuyas tesis serían medulares al cognitivismo:

> [The] C[omputational] T[heory of] M[ind]'s claim, after all, is not that production-model computers provide a good metaphor for the mind, but that the exact mathematical notion of computation provides the right sort of resources for supplementing a representational account of intentionality with a computational account of cognitive processes.

Pero, naturalmente, Pylyshyn es consciente de que la interpretación de la homología entre cognición y computación como una relación literal de instancia a clase no es hegemónica ni siquiera en los propios predios de la psicología cognitiva. De hecho, es posible diferenciar en el seno del cognitivismo entre quienes adoptan la lectura literal propugnada por Pylyshyn y quienes prefieren quedarse con una metáfora computacional entendida como genuina metáfora. Exactamente en esos términos distingue Rivière (1991b: 140) entre un "paradigma computacional-representacional", que haría hincapié en la justificación de los constructos teóricos del cognitivismo mediante su imbricación en algoritmos de cómputo, y un "enfoque de procesamiento de la información", que se conforma con dar su teorización por vagamente justificada, con carácter general, por la analogía entre mentes y programas, pero reserva la justificación de conceptos teóricos particulares a la evidencia observable y, sobre todo, rehúye de lecturas ontológicas, literales, de la metáfora. Se trata, en suma, la distinción entre una concepción fuerte y una concepción débil de la investigación en inteligencia artificial en la que tanto ha insistido Searle (1980, 1992), y que ha permitido a muchos psicólogos adoptar posiciones que acaso encontraran menos comprometedoras. Como el propio Pylyshyn anota:

> Although "computer simulation" has been a source of much interest to psychologists for several decades now, it has frequently been viewed as either a useful metaphor or as a calculation device, a way to exhibit the consistency and completeness of independently framed theories in some domain of cognition. The very term *computer simulation* suggests verisimilitude –imitation– rather than a serious proposal as to how things really are. (Pylyshyn 1984: *xiii-xiv*)

Su diagnóstico, sin embargo, es que esa tendencia a esquivar la lectura literal de la homología proviene de una deficiente comprensión del concepto de computación:

> It is the failure to distinguish computation as a type of process from the particular physical form that it takes in current computing machines that has prevented many people from taking computation as a literal account of mental process. If we understand computation at a fairly general level (as, in fact, it is understood in theoretical computer science), we can see that the idea that mental processing is computation is indeed a serious empirical hypothesis rather than a metaphor. (Pylyshyn 1984: 55)

No sería particularmente arduo añadir a ese diagnóstico consideraciones más o menos retóricas sobre nuestra renuencia a dar por buena ésta o cualquier otra tesis de la que –temamos– pudiera desprenderse una degradación de alguna particular idea de la dignidad humana: el cognitivismo se enfrentaría entonces a los mismos, espurios, motivos de repulsa que en su día acosaron al heliocentrismo, al evolucionismo o, dicho sea de paso, al psicoanálisis y al conductismo, según sus propios adalides. Anécdotas no faltarían: el joven Vaucanson tuvo que ver como sus primeros ingenios eran destruidos por mandato del superior de la Orden de los Mínimos de Lyon; el reputado relojero Pierre Jaquet-Droz, autor de primorosos mecanismos que remedaban los movimientos de un niño que dibujaba sobre un pupitre, una pianista y un escritor, viajó a Madrid en 1758 y, ante la fascinación de Fernando VI y el espanto de la corte, se vio obligado a mostrar al arzobispo Manuel Quintano, confesor del rey e inquisidor general, que no había brujería en aquellos artilugios –algunos relatos, probablemente legendarios, aseguran que fue hecho preso a instancias del Santo Oficio y se confiscó su metálica progenie (*cf.* por ejemplo Bueno y Peirano 2009: 27, Nocks 2007: 33-34); de lo que no hay duda es de que el reloj de Jaquet-Droz, con su flautista que toca una melodía al dar las horas, todavía decora el Palacio Real. Se diría, sin embargo, que no vale la pena el intento de detallar ese capítulo de la historia de la superstición, toda vez que cuanto hay de falaz en las argumentaciones fundadas sobre tales recuentos ha sido severamente desembozado por Searle (1992: 5) al describir "[…] la maniobra de la edad-heroica-de-la-ciencia" como cualquier intento de hacer ver que el mero hecho de que nuestro interlocutor no convenga en determinada tesis lo convierte en el cardenal Belarmino –y a uno mismo, claro, en Galileo.

Sea como sea, eludir el corolario de la concepción funcionalista de los estados mentales que cristaliza en la posición de Phylyshyn no se justifica porque lo anhelemos: exige una argumentación cuidadosa, tanto más cuanto el territorio que queda entre el funcionalismo y las cercanas comarcas del conductismo lógico y el fisicalismo de tipos es –hasta donde sabemos– exiguo. Dicho de otro modo: es difícil rebatir el compromiso funcionalista con la identidad literal entre (tipos de) estados mentales y (tipos de) estados computacionales sin que ello entrañe volver al compromiso con la identidad literal entre (tipos de) estados mentales y (tipos de) estados físicos, o de conjuntos de disposiciones –y con ello a sufrir, acaso

recrudecidas, las objeciones que arruinaron estas teorías de lo mental. De nada sirve, en cualquier caso, el voluntarioso rechazo de la conclusión indeseada –el trámite que aplica Gardner (1985) cuando anota:

> Si bien el nexo entre la computación y el cognitivismo es contingente y no forzoso, lo cierto es que el destino de la ciencia cognitiva ha quedado íntimamente ligado al de la computadora. (Gardner 1985: 412)

Pero para poder sostener razonadamente que el cognitivismo y la noción de computación están ligados de modo meramente casual, debido –digamos– a las contingencias históricas que confluyeron sobre la comunidad científica anglosajona hacia el final de la Segunda Guerra Mundial, sería preciso conformarnos con una definición de cognitivismo tan vaga que quedaría satisfecha prácticamente por cualquier variedad de psicología científica distinta del conductismo –y, con ello, el propio argumento de la contigencia histórica perdería su sentido. Que las tesis funcionalistas surgieran en parte del estudio de la teoría de autómatas es sin duda una contigencia histórica: bien podían haber surgido únicamente, por ejemplo, del intento de afrontar conjuntamente las limitaciones de la tesis de identidad psicofísica y del conductismo lógico, o de la reflexión sobre las consecuencias de la reflexión wittgensteiniana acerca del significado para la semántica de las actitudes proposicionales, o de innumerables otros manantiales. Aun así, no es descabellado sostener que las tesis funcionalistas, salvo que se acoten convenientemente, implican la identidad entre estados mentales y estados computacionales, y la implicarían aunque la noción de computación no hubiera sido forjada todavía, ni las computadoras hubieran comenzado a fabricarse[135]. Por mucho que una voz tan autorizada como la de Ulric Neisser desestimara en 1967[136] los modelos computacionales del pensamiento que entonces estaban sobre la mesa, tildándolos de "simplistas" e "insatisfactorios desde un punto de vista psicológico", e insistiera en el hecho difícilmente cuestionable de que "[…] el ocasional uso analógico de conceptos tomados del ámbito de la programación informática no implica un compromiso con la simulación computacional de procesos psicológicos" (Neisser 1967: 9), lo cierto es

[135] Una aseveración parecida a la de Gardner formula von Wright (1971: 9) respecto de la relación entre el vertiginoso desarrollo de la lógica formal en las primeras décadas del s. XX y el reverdecer del positivismo –precisamente en su encarnación conocida como positivismo *lógico*– durante el período de entreguerras, relación que tilda de "accidente histórico". También Leahey (2005: 395) presenta un razonamiento similar, en este caso acerca de la relación entre los adelantos de la teoría de la computación –que liga al desarrollo de la investigación en inteligencia artificial– y las teorías del procesamiento de la información: "[…] aunque las teorías del procesamiento de la información eran independientes de las teorías computacionales de la I[nteligencia] A[rtificial], de alguna manera eran tributarias de ellas […]". Que salvedades parecidas a las que pueden plantearse a la tesis de Gardner puedan también oponerse a las de von Wright o Leahey es una perspectiva que no cabe escrutar aquí. Dado que en ambos casos son distintos los *relata*, también puede serlo, qué duda cabe, el valor de verdad de la afirmación.

[136] En la misma obra en la que a todas luces parece Neisser haber inaugurado el uso actual de la expresión "psicología cognitiva", que le daba título.

que, llegada la hora de defender a la psicología de la objeción de que ésta no es más que una especie de divertimento "[…] mientras llega el bioquímico […]", el propio Neisser (1967: 6) recurre al "[…] familiar paralelismo entre hombre y computadora […]" para señalar, como el más aplicado funcionalista, que el trabajo del psicólogo es análogo al del ingeniero que trata de desentrañar mediante qué rutinas un determinado programa realiza tal o cual tarea. La clave de la analogía, según el propio Neisser (1967: 8) no reside sino en la constatación, que podrían haber suscrito Putnam o Fodor de que:

> Although a program is nothing but a flow of symbols, it has reality enough to control the operation of very tangible machinery that executes very physical operations.

Es el "paralelismo entre hombre y computadora", pues, en el pensamiento de Neisser, lo que pese a su rechazo declarado de la simulación computacional da respaldo a la afirmación de que "[…l]a psicología trata acerca de la organización y uso de información, no acerca de su representación en el tejido orgánico" (Neisser 1967: 281)[137]. Dicho de otro modo: esas tangibles máquinas, tan rudamente materiales –o mejor, los programas que las gobernaban, tan, si se quiere, rudamente inmateriales–, proporcionaban una forma rotundísima de exorcizar "[…] esos peculiares y fantasmáticos miembros de su cohorte" –la de la mente– "(los pensamientos, los recuerdos, las imágenes, los recuerdos y creencias) para ser readmitidos en el recinto respetable de la ciencia" (Rivière 1991b: 135-136). O, como con meridiana claridad lo ha expresado Johnson-Laird –en unas líneas que también Rivière cita *in extenso*:

> […] la invención del ordenador digital […] ha obligado a la gente a pensar de una forma nueva sobre la mente. Antes de la computación había una distinción clara entre cerebro y mente; uno era un órgano físico y la otra una "no entidad" fantasmática que difícilmente resultaba un tema de investigación respetable. (Se consentía que los adultos pudieran hablar de ella en privado, siempre y cuando comprendieran que, en realidad, no existía.) Después de la llegada de los ordenadores no cabe semejante escepticismo: una máquina puede controlarse mediante un "programa" de instrucciones simbólicas, y no hay nada de fantasmal en un programa de ordenador. Quizá, y en gran medida, la mente es para el cerebro lo que el programa es para el ordenador. De esta manera, puede haber una ciencia de la mente. (Johnson-Laird 1988: 13-14)

No es sólo, aunque es muy decisivamente, de esa legitimidad digamos fundacional de lo que la reflexión sobre las máquinas de cómputo podía dotar a una rehabilitación de la psicología de los procesos internos: es, también, del ejemplo nítidisimo de cómo la minuciosa formalización de cada uno de los pasos que forman un proceso tan sumamente complejo como los que despiertan el interés del psicólogo

[137] De hecho, Neisser (1967: 6) recurre incluso a la misma analogía entre psicología y economía que encontraremos luego en Fodor (1974: 103-104), haciendo ver la irrelevancia de las múltiples y heterogéneas implementaciones físicas que pueden darse al dinero de cara a buena parte de los objetivos de la investigación en economía.

puede constituir una explicación de dicho proceso. Los algoritmos y programas de las computadoras, así, comenzaban a perfilarse como parangón de las teorías psicológicas. Como apunta Hatfield:

> The symbolist view was introduced into cognitive psychology and cognitive science as a result of taking the computer metaphor very seriously. It made sense to do this because perception and other cognitive processes are quite complicated, and the computer provided a powerful example of a device that could perform complicated information-handling tasks and whose operation was understood in detail. By treating the perceptual [… (or the cognitive)] system as a computational device, investigators hoped to be able to capture its complexity within a well-understood type of model. (Hatfield 1989: 264-265)

Ya en 1960, no bien incorporaron el concepto de *plan* –"[…] *cualquier proceso jerárquico del organismo que puede controlar el orden en que se tiene que realizar una secuencia de operaciones*" (Miller, Galanter y Pribram 1960: 26-27)– a su proyecto de descripción de la estructura de la conducta, Miller, Galanter y Pribram (1960: 27) anotaban sin titubeos que "[…p]ara un organismo, un *plan* es esencialmente lo mismo que un programa es para un ordenador". Pero si los programas de ordenador equivalen de forma genérica a los planes que rigen la ejecución de operaciones de los organismos, basta suponer que un programa simula con éxito un proceso psicológico determinado para que tengamos al alcance de la mano la conclusión de que el programa equivale "esencialmente" al proceso –cognición *es* computación. Asumido que el programa tendrá presumiblemente, al menos para sus programadores, una estructura transparente, viene dada además la conclusión de que dicho programa constituye una explicación legítima del proceso al que equivale: su teoría, o acaso un modelo que la encarna. Extender la equivalencia entre programas y planes hasta hacerla abarcar cualquier proceso psicológico no es difícil, dada la amplitud de la noción de programa implicada: los propios Miller, Galanter y Pribram (1960: 234) consignarían ésta como una de las conclusiones últimas de su trabajo al anotar que "[...e]l programa de ordenador que imita un proceso está llegando a ser una teoría de ese proceso tan aceptable como sería la ecuación que lo describe".

En su brillante estudio de la historia de la simulación mecánica de la conducta de los organismos durante el s. XX, antes de la consagración como disciplinas de la cibernética y la inteligencia artificial, ha recabado concienzudamente Cordeschi (2002) un buen número de ejemplos que muestran cómo esta homología entre programa y teoría se había ido forjando de forma paulatina en la reflexión de los ingenieros, fisiólogos y psicólogos que compartían aquel proyecto. La idea se atisba en el trabajo de ingenieros como S. Bent Russell ya en 1913: el propósito de la máquina hidráulica que éste describiera no era otro que "encarnar" ciertas hipótesis comúnmente aceptadas sobre el flujo de la señal nerviosa y así poder "[…] comparar los resultados obtenidos en la máquina con los ofrecidos por conexiones nerviosas vivas" (Bent Russell 1913 *apud* Cordeschi 2002: 31); hemos visto ya que Meyer (1913), en el contexto de su prolongada disputa contra el vitalismo de McDougall, citaría el trabajo de Bent Russell como prueba de que una máquina que no habita fantasma

alguno es capaz de mostrar formas elementales de aprendizaje. En el extraño artefacto diseñado en 1912 por dos ingenieros expertos en radiocontrol, John Hammond Jr. y Benjamin Miessner –un "perro eléctrico" que podía ajustarse para que siguiera la luz de una linterna o huyera de ella, y que ya había despertado el interés del estamento militar (*supra*)– veía Jacques Loeb (1918: 69 *apud* Cordeschi 2002: 7) un claro refrendo de su propia teoría del fototropismo, "[…] puesto que esta teoría sirvió de base para la construcción de la máquina".

La construcción de modelos mecánicos inspirados en las hipótesis de una teoría determinada abría, a ojos de John M. Stephens (1929), un camino nuevo para la verificación y la construcción de teorías, y constituía en ese sentido un auténtico descubrimiento de orden epistemológico. Las primeras líneas de su trabajo sobre "Una explicación mecánica de la ley del efecto" son reveladoras: "As this paper attempts to present both the description of a learning machine and a theory of learning, it might be helpful to insert a statement about the inter-relations between the apparatus and the theory"; la clave de esa declaración llega a renglón seguido: "As a rule, when analysis has reached a certain point, synthesis becomes possible" (Stephens 1929: 442). Los estudios del joven Stephens sobre Thorndike despertaron el interés ni más ni menos que de Hull, quien lo invitó a presentarlos en la Conferencia de la *Eastern Branch* de la *American Psychological Association* en 1930 (Evans 2000: 15); el propio Hull, que no en vano había estudiado lógica e ingeniería además de psicología, profundizaría como hemos visto antes que nadie en las perspectivas metodológicas que abría la simulación. Decidido a replicar mecánicamente el aprendizaje por ensayo y error remedando también las limitaciones y errores que se advertían en los sujetos experimentales estudiados por Hull, Douglas G. Ellson (1935: 94 *apud* Cordeschi 2002: 94) apostaba firmemente por tomar aquellos artefactos como "hipótesis mecánicas". También Tolman (1939) describiría su "cochinilla esquemática" como una *ejemplificación* de la teoría del aprendizaje vicario por ensayo y error (*cf*. Cordeschi 2002: 137). En 1943, K.J.W. Craik –a cuyo papel en la relativa refractariedad de la comunidad psicológica británica al conductismo se ha aludido *supra*– podía ya trazar un distinción entre un mecanicismo digamos filosófico, que se limitaba a aseverar el carácter mecánico de los procesos biológicos o psicológicos, y un mecanicismo que cabría describir como "científico" en tanto en cuanto la teoría se toma como "[…] a hypothesis which should, if followed out, indicate how and where [the theory] […] breaks down" y, por tanto, nos compromete a formular "[…] a definite plan of a mechanism which would fulfill the requirements" (Craik 1943: 52 *apud* Cordeschi 2002: *xvi*). El hito cardinal de este nuevo mecanicismo era, para Craik, el trabajo de Hull.

Mediado el siglo, así pues, el clima intelectual en ciertos círculos era ya bien propicio para que el vínculo entre simulación y teorización fuera tornándose más patente. Lo que Wallace (1952) subrayaba, en efecto, en relación con el pequeño vehículo al que había dotado de la capacidad de hallar y recordar la salida de un laberinto, era que el diseño de aquel ingenio constituía una contribución a la explicación del aprendizaje (Wallace 1952: 121 *apud* Cordeschi 2002: 163). Sólo un año

después, por cierto, J.A. Deutsch anunciaría –recordemos– que su invento –también construido para aprender a recorrer laberintos– perfilaba "[…u]na nueva clase de teoría de la conducta" (Deutsch 1953, *supra*), si bien pronto matizaría la idea al apuntar que la máquina debía considerarse un modelo que encarnaba la teoría (Deutsch 1954, *supra*). Para L.B. Wyckoff (1954), su modelo electrónico era un intento de cuantificar una posible ley del aprendizaje que Skinner había formulado en términos cualitativos –a saber: que los estímulos discriminativos operan como reforzadores secundarios– y, al hacerlo, "[…] revelar las deficiencias de la teoría" (Wyckoff 1954: 95 *apud* Cordeschi 2002: 170); se trata, pues, de un instrumento para el sometimiento de la teoría a prueba empírica y, en consecuencia, para afinarla. Lo que instigaba el interés del propio W. Grey Walter (1950, 1951, *supra*) por las célebres "tortugas" que, sin abandonar las investigaciones sobre electrofisiología que habían forjado su reputación, venía construyendo desde finales de la década de 1940 –como la *Machina speculatrix*, capaz de explorar su entorno, seguir una fuente de iluminación manteniéndose a cierta distancia de ella, esquivar obstáculos en su camino, y alternar entre una y otra cuando varias luces estaban presentes, o la *Machina docilis*, que además podía aprender– residía, como también recuerda Cordeschi (2002: 166), en el hecho de que Elmer y Elsie –las tortugas– encarnaban "[…] una hipótesis teórica". Para MacKay (1954: 404), un modelo mecánico nos proveía de "[…] a kind of template which we construct on some hypothetical principle, and then hold up against the real thing"; al comparar la plantilla con el fenómeno estudiado se harían patentes las eventuales discrepancias (*cf*. Cordeschi 2002: 248).

Por las mismas fechas, H.E. Coburn (1951, 1952, 1953a, 1953b) presentaba –recordemos– un modelo cuantitativo de la función cerebral –no, *nota bene*, de las estructuras nerviosas– que estaba llamado a conformar a un tiempo un croquis para la construcción de máquinas inteligentes, una herramienta de investigación neuropsicológica, *y una teoría* de la conducta (*cf*. Cordeschi 2002: 166-167). La misma osadía puede verse en Broadbent (1957), quien haciéndose eco de los trabajos de Deutsch caracterizó los modelos mecánicos como teorías "[…] expresadas en componentes materiales en lugar de en símbolos abstractos, como palabras o expresiones matemáticas" (Broadbent 1957: 205 *apud* Cordeschi 2002: 169). La diferencia albergaba un buen augurio: lo había adivinado Grey Walter (1957: 8, *cf*. Cordeschi 2002: *xvi*, que no ubica la cita), tras acuñar para los modelos mecánicos la feliz pero fallida denominación de "hipótesis cristalizadas", cuando apuntaba que

> […i]f in the testing [working models] […] fall short of expectation or reality, they do so without equivocation. The model hypothesis cannot bend or flow –it breaks with a loud crack– and from the pieces one can build a better model.

La misma promesa, dejando entrever la misma idea de –por así decir– la *sinceridad* de los modelos artificiales, quedó consignada al escribir Coburn (1952: 458 *apud* Cordeschi 2002: 170) que la expresión verbal de las teorías tendía a ocultar sus

inconsistencias, mientras que su realización en estructuras físicas las dejaba rápidamente al descubierto. Así, lo que Wyckoff veía como una herramienta útil para la construcción de teorías se perfila ya en Coburn, y luego en Broadbent, como una expresión distinta, pero no menos lícita, de la propia teoría[138].

También Newell, Shaw y Simon (1958), en un trabajo con el vigorosísimo carácter germinal de "Elements of a Theory of Human Problem-Solving", afirmaban con rotundidad que "[…] an explanation of an observed behavior of an organism is provided by a program of primitive information processes that generates this behavior" (Newell, Shaw y Simon 1958: 151), si bien al instante matizaban que un programa, "[…] visto como una teoría", es extremadamente específico, pues se ciñe a un organismo y una situación en particular. Ya en 1956, Simon y Newell habían comparado los pros y contras de la formulación verbal y la formulación matemática de una teoría con los de su formulación a modo de programa informático; la tesis de que el programa es de hecho una formulación de la teoría se reiteraría en Newell y Simon (1963). Las primeras líneas de "An Information Processing Theory of Verbal Learning", el célebre informe P-1817 de la División de Matemáticas de Rand Corporation donde Edward A. Feigenbaum (1959) presentó su *Elementary Perceiver and Memorizer*, EPAM, proclamaban con rotundidad que "EPAM is a theory of human verbal learning, expressed in a language for a digital computer, IPL-V" –a renglón seguido Feigenbaum procedía a reconocer su deuda intelectual con Newell y Simon.

La cuestión iría poco a poco envolviéndose en cierta controversia: Dennis Gabor, un físico húngaro que había adquirido ya cierta celebridad por sus trabajos sobre la reconstrucción del frente de onda, u holografía, rechazaría de forma un tanto expeditiva las pretensiones teóricas de los modelos mecánicos del aprendizaje (Gabor 1956 *apud* Cordeschi 2002: 163); Andrew Gordon Pask (1961), un ingeniero de minas que por entonces comenzaba en Londres su estudios doctorales en psicología, circunscribiría en cambio el valor explicativo de los modelos mecánicos a los casos en los que éstos no se limitaran a imitar las conductas de un organismo, sino que encarnasen los mismos *principios funcionales* que éste –un movimiento parejo al que en Fodor (1968: 177) conduciría a la idea de equivalencia funcional fuerte (*cf.* Pylyshyn 1984: *xv*, *infra*). Tiempo atrás, en la pleamar del conductismo, la misma idea se había podido atisbar en los trabajos de Ross (1933, 1935, 1938), para quien los modelos mecánicos se perfilaban como herramientas de contrastación de las teorías psicológicas. Al desplegar las condiciones bajo las que tal contrastación habría de

[138] Conviene señalar que la diferencia entre que un modelo mecánico sirva como herramienta en la verificación o en la construcción de teorías y que se identifique con la propia teoría, o con una expresión suya, es pasada por alto en el inestimable recuento elaborado por Cordeschi (2002), donde ambas posiciones se tratan indistintamente; sí se distingue, en cambio, la idea de que los modelos mecánicos atesoren un valor explicativo de la de que sean meras pruebas de posibilidad del carácter natural de determinados fenómenos. La formulación preferida por Cordeschi pasa por afirmar que el modelo *encarna* de forma operativa las hipótesis que constituyen la teoría (*cf.*, por ejemplo, Cordeschi 2002: 33).

tener lugar, el margen de independencia entre función y estructura que vertebraría las tesis funcionalistas quedaba ya cristalinamente expuesto:

> […I]t may be possible to test the various psychological hypotheses as to the nature of thought by constructing machines in accord with the principles that these hypotheses involve and comparing the behavior of the machines with that of intelligent creatures. Clearly, this synthetic method is not intended to give any indication as to the nature of the mechanical structures of physical functions of the brain itself, but only to determine as closely as may be the type of function that may take place between "stimulus" and "response" as observed in the psychological laboratory and in ordinary uncontrolled learning and thinking. Only analogies which will work are sought, not imitations of nerve, brain and muscle structure. (Ross 1935: 387 *apud* Cordeschi 2002: 109)[139]

Es precisamente en un estudio sobre *La explicación psicológica* donde toma cuerpo la reflexión en torno a "la lógica de la simulación" (Fodor 1968: 159) que daría buena parte de su impulso a la idea de que existe un estrecha relación entre la teoría que explica una conducta y el modelo computacional que la remeda: "[…] se llega a decir" –da fe de ello Fodor (1968: 159) en las primeras líneas del capítulo que cierra su investigación– "que comprender las operaciones de un computador que sea capaz de simular una forma determinada de conducta es tanto como comprender la conducta misma". En el intento de delimitar las condiciones bajo las que una afirmación como esa pueda ser verdadera se hacen imprescindibles tanto la distinción entre la conducta efectivamente exhibida por un organismo –o una máquina– en un conjunto cualquiera de situaciones (su *actuación*) y las conductas que podría exhibir en condiciones contrafácticas (su *competencia*), como la idea de un criterio de equivalencia más severo que la mera indiscernibilidad conductual sobre la que Turing (1950, *supra*) había hecho reposar su juego de imitación. En efecto, que simular una conducta valga tanto como explicarla exige en primer lugar –ése es el dictamen de Fodor– que lo simulado abarque la totalidad del repertorio conductual del organismo: el repertorio conductual, y no la actuación del organismo en tales o cuales circunstancias, se convierte en "[…] el objeto primario de la simulación" (Fodor 1968: 170). Pero además, no podremos conceder carácter explicativo a una simulación mientras no constatemos que "[…] los procesos internos de la máquina sean similares, en sentido relevante, a los del organismo que [… la máquina]

[139] Es interesante constatar como la misma idea de la relación entre modelos mecánicos –o programas– y teorías se da en Rochester *et al*. (1956), si bien completamente desligada del aire funcionalista que se advierte en Ross (1935): lo que el equipo de Rochester trata de simular, de hecho, son los patrones de actividad de grupos de neuronas que, de acuerdo con la teoría de Hebb (1949), subyacerían al aprendizaje. En el mismo número de las *Transactions of Information Theory*, que editaba entonces el *Institute of Radio Engineers* de Nueva York, en el que se publicó el artículo de Rochester y su equipo aparecía también el trabajo de Chomsky (1956) en torno a "Tres Modelos para la Descripción del Lenguaje" (*supra*), donde alzaba el vuelo el estudio de las propiedades abstractas de diferentes tipos de gramática y los recursos necesarios para su implementación mecánica. La convivencia entre enfoques muy dispares en su adhesión a la estructura neurológica nunca quedaría, como veremos, del todo arrinconada en el seno del cognitivismo.

pretende simular" (Fodor 1968: 176) –es decir, que entre la máquina y el organismo se dé no sólo la relación de equivalencia débil definida por el test de Turing, sino una de equivalencia fuerte. Naturalmente, la primera exigencia –que la máquina simule la conducta del organismo también en condiciones contrafácticas– nos conduce de lleno a advertir que la evidencia que pueda obrar en favor de un determinado intento de simulación es de naturaleza inductiva, y se ve aquejada de la misma debilidad que en términos generales afecta al razonamiento inductivo: "[…] ningún número de simulaciones satisfactorias puede servir de base *lógicamente* suficiente para determinar el valor explicativo de un modelo psicológico" (Fodor 1968: 172). Del mismo modo, la segunda exigencia –que la máquina simule los procesos internos del organismo además de su conducta observable– nos aboca al espinoso problema de hallar "[…] un criterio para determinar cuándo habría que considerar a las operaciones de aquélla y de éste como semejantes en un sentido relevante" (Fodor 1968: 180)[140]. El caso es que, mal que nos pese, no hay otra forma de resolver ese problema que el propio desarrollo de la investigación psicológica:

> Las teorías psicológicas tienen precisamente como objetivo el explicitar el grado y la naturaleza de esas interconexiones que se dan entre los procesos psicológicos. El asunto está en que, al hacer eso, quedan implícitamente determinadas las condiciones requeridas para la equivalencia funcional de los estados y procesos postulados por la teoría. No es fácil, desde luego, ver cómo podrán efectuarse tales determinaciones previamente a la construcción de la teoría psicológica adecuada. (Fodor 1968: 183)

Ninguno de estos escollos, con todo lo escabrosos que pueda antojársenos, despunta sólo en los mapas de la simulación computacional de procesos psicológicos; antes al contrario, conforman dificultades bien conocidas en cualquier ámbito de la investigación científica. Que el tránsito de la generalización al vigor nomológico se despliega en el proceloso terreno de lo contrafáctico, y con ello de la inducción, que el modo de tejer un vocabulario teórico condiciona el conjunto de generalizaciones que nos será dado atrapar en sus redes –más aún: que demarca de hecho el objeto de la investigación (*cf*. Pylyshyn 1984: 271, *infra*)–, que las decisiones cruciales no pueden tomarse *a priori*, sino que deben aguardar al paso de los hallazgos –o los desengaños– empíricos: cualquiera de esos rasgos es tan característico de la simulación cognitiva como de la física; reviste, en fin, "[…] un carácter

[140] Problema que, dicho sea de paso, surge igualmente al tratar de fijar razonadamente un criterio de equivalencia débil, conductual: determinar a qué tipo de conducta pertenece una conducta concreta de la máquina o del organismo, como paso previo para establecer si ambas pertenecen al mismo tipo, requiere haber delimitado "[…] un vocabulario teorético en el que habrá que describir las conductas relevantes". Ahora bien, determinar el vocabulario teórico –y con ello, el nivel de descripción– idóneo no es algo que quepa estipular, pues "[…n]o toda elección de niveles conduce a una representación sistemática del repertorio conductual del organismo" (Fodor 1968: 179) –se perfila ya aquí el vínculo entre vocabulario teórico y capacidad de captura de generalizaciones que veremos desplegarse en Fodor (1974: 101) y en Pylyshyn (1984: 16-17), *infra*. Una resolución estipulativa del asunto es, precisamente, la que ensaya Turing (1950), al dictaminar las condiciones en las que se desarrolla el juego de imitación.

completamente general en cuanto a su aplicabilidad a las teorías científicas" (Fodor 1968: 173). La lección crítica de las reflexiones de Fodor es, así pues, que

> […] la cuestión "¿Qué relación existe entre simular una conducta y explicarla?" no es más que un caso particular de la cuestión general que se plantea en torno a las teorías científicas –a saber, "¿Qué relación existe entre que una teoría sea compatible con los datos experimentales y que sea verdadera?". (Fodor 1968: 173)

No es difícil ver que una formulación bastante precisa de la analogía entre programas y teorías está implícita en esa conclusión: un programa que logra simular tal o cual conducta es capaz de explicarla en la misma medida y bajo las mismas restricciones que es verdadera una teoría compatible con los datos experimentales concernientes a dicha conducta. Con algo más de laxitud apunta Fodor poco antes que las cláusulas que ha venido describiendo deben verificarse "[…] si es que queremos que el programa de la máquina sea considerado como una adecuada teoría psicológica del organismo" (Fodor 1968: 172), pero también, después, que para establecer que esas cláusulas se cumplen es preciso "[…] determinar hasta qué punto la teoría psicológica puesta en juego por el programa de la máquina constituye una explicación simple y adecuada de la conducta del organismo" (Fodor 1968: 189), o, dicho de otra manera, "[…] si la teoría psicológica que es llevada a cabo por el programa de la máquina es verdadera al aplicarla al organismo" (Fodor 1968: 191). Las teorías son, así pues, *puestas en juego* y *llevadas a cabo* por programas; los programas, al mismo tiempo, *son* teorías.

Un año antes de la publicación de la monografía de Fodor, un breve artículo de Nico Frijda (1967) había señalado que, si bien un programa puede servir no sólo como "[…] un medio para poner a prueba la consistencia y la suficiencia de una teoría", desvelando sus implicaciones y presupuestos, y como un procedimiento heurístico en la búsqueda de nuevas hipótesis, sino también como una "[…] formulación de la teoría" (Frijda 1967: 59) despojada de toda ambigüedad, es importante no consentir la inexacta afirmación de que un programa *es* una teoría:

> Rather, a program *represents* a theory. It does this with the help of a series of mechanisms which are irrelevant to the theory or which the theory might explicitly disclaim. The lower order subroutines and a number of technical necessities are determined by the particularities of the programming language, the mode of operation of the particular computer, and the special limitations inherent in serially operating digital machines. Many operations, too, are just shortcuts for convenience or results of ignorance about the psychological mechanisms involved. (Frijda 1967: 60)

La insistencia de Frijda en que "[…] la estructura del programa refleje la estructura de la teoría", de forma que las rutinas teóricamente relevantes queden aisladas de las operaciones auxiliares y los detalles técnicos (Frijda 1967: 61), entronca con su preocupación por el hecho de que en las tareas de simulación el incumplimiento de esos requisitos "[…] seriously detracts from the value of the work", dado que "[…] performance as such is not so important here as convincing the reader that the

reasons for this performance are plausible" (Frijda 1967: 62)[141]. Se trata, en suma, de los mismos escrúpulos respecto a la verosimilitud psicológica de la simulación que hemos entrevisto en Pask, y que tomarían cuerpo en Fodor por medio de la noción de equivalencia fuerte.

Con más concisión que moderación en la generalidad de la tesis registraría años después Cummins (1986: 125) que "[...] one of the important insights of the cognitive science movement is that a program is a theory". Algo más precavido, Thagard (2005: 32), en la estela de Fridja (1967), insiste en diferenciar la teoría cognitiva, que "[…] postula una serie de estructuras representacionales y una serie de procesos que operan sobre ellas", el modelo computacional, que "[…] vuelve más precisos esos procesos y estructuras a través de interpretaciones que consisten en analogías con programas de computación formados por estructuras de datos y algoritmos", el "[…] programa de *software* creado con un lenguaje de programación" y, por último, la plataforma de *hardware* en la que se ejecuta el programa. También Johnson-Laird (1983) prefería discernir entre su teoría de modelos mentales y los diversos programas en los que ésta se despliega, que encarnan determinadas hipótesis de la teoría y establecen con ella un "proceso dialéctico" (Johnson-Laird 1983: 13).

Sea como sea, la equivalencia entre programas y teorías, entendida de forma más o menos laxa, ha venido siendo sin lugar a dudas uno de los motores de la investigación cognitivista, por razón de los anchísimos horizontes metodológicos que abre. Es más, es en buena medida el conjunto de hábitos intelectuales instigados por el empleo de programas como trasuntos de nuestras teorías lo que –como acierta en concluir Cordeschi (2002: *xviii*)– nos ha forzado a preguntarnos "[…] what the *right* level of abstraction for constructing explanatory models of mental life might be", propiciando así que quedaran radicalmente sometidas a reconsideración "[…] the customary taxonomies of the sciences of the mind".

Ahora bien: la idea de que un programa pueda constituir en sí mismo una cierta expresión de una teoría –con el mismo derecho, digamos, a ser identificado con la propia teoría que el que pudiera corresponder a su expresión verbal o matemática– no casa del todo limpiamente con la lectura literal de la relación entre cognición y computación –la tesis de que los procesos cognitivos son procesos computacionales, o, si se quiere de que mentes y programas pertenecen a una misma clase natural, como lo hacen cerebros y computadoras. Las aristas son palpables: de una misma entidad, el programa, parece requerirse que sea idéntico a dos distintas, el proceso psicológico –o, en general, la mente– y la teoría que lo explica. Leves refinamientos conceptuales deberían bastar, de todas formas, para limar asperezas tan leves, pues no es mucho pedir que distingamos entre la ejecución de un programa –o, si se

[141] Por otra parte, la necesidad de distinguir entre programas discernibles al pie de la letra pero que, en tanto en cuanto su discernibilidad no reside en aspectos relevantes desde la perspectiva teórica, constituyen formulaciones de la misma teoría bien puede contarse entre las preocupaciones que acabarían dando forma a la distinción entre el nivel algorítmico y el nivel computacional que encontraremos nítidamente expresada en Marr (1982, *infra*).

quiere, el programa en ejecución– y el programa ejecutado, asimilando aquel al propio proceso cognitivo que el programa pretende replicar y éste a la teoría que explicaría dicho proceso –o, si se prefiere, a una expresión entre otras de esa teoría. La idea de que el programa constituye una formulación operativa de la teoría, o, más modestamente, de algunas de sus hipótesis, encontraría así –digamos– su nicho ontológico: decimos del programa, en efecto, que es una formulación operativa de la teoría en tanto se diferencia de otras formulaciones por su potencialidad de obrar, es decir, de erigirse en una instancia del propio fenómeno que explica.

Al margen de esas acaso menudencias analíticas, lo cierto es que las historias de una y otra idea –que el programa es el *explanandum*, que en realidad es el *explanan*s– se entrelazan como ramas de una enredadera. Igual que sucede con la intuición de que un modelo mecánico puede llegar a ser en cierto modo una explicación de aquello que remeda, la tesis de que la relación entre ciertos procesos psicológicos y sus réplicas artificiales pudiera ser más estrecha que la que una mera cercanía metafórica requiere había ido haciéndose hueco poco a poco entre quienes consagraban su esfuerzo a la construcción de tales réplicas, o quienes ponían cuidado en reflexionar sobre ellas –antes de Pylyshyn (1984), de Miller (1984) e incluso de Newell y Simon (1972). Es significativo, además, que las ligaduras entre ambas ideas conduzcan una y otra vez a la misma constatación: tanto si hemos de conceder a un autómata virtudes explicativas respecto a aquello que hemos logrado que calque, como, con más razón, si hemos de considerarlo sin ambages como un integrante de la misma clase natural a la que pertenece lo que los croquis del autómata convenimos en que explican, habremos de exigir al menos que el modo en que el autómata lleva a cabo sus quehaceres sea significativamente parecido al modo en que estos se despliegan en el caso de aquello que el autómata, al mismo tiempo, explica y constituye. No es difícil anticipar, así pues, que la cuestión de qué debamos entender por un parecido significativo acabará por resultar decisiva –o, dicho de otra forma, que será crucial elucidar bajo qué nivel de abstracción deberán diseñarse los modelos y, en consecuencia, articularse el vocabulario explicativo de las teorías.

Así, en un tempranísimo cotejo de diversos intentos de zanjar la controversia sobre el vitalismo remedando mecánicamente fenómenos propios del ámbito orgánico, como el comportamiento de una ameba –la misma controversia que, como se ha visto, enfrentara a Meyer (1913) y McDougall (1911), y en la que el propio Jennings (1910) tomaría partido frente a Jacques Loeb (1900) a cuenta de la naturaleza de los tropismos–, Jennings (1904) hacía valer un sensato *caveat*: habría que dilucidar si la simulación del comportamiento de un ser vivo por medio de un mecanismo inorgánico compartía con dicho comportamiento su raigambre causal, o más bien constituía una simple imitación. Es ese esfuerzo por desgranar las "condiciones paralelas" (Jennings 1910: 361) que en un organismo y en una máquina pudieran dar lugar a respuestas genuinamente equiparables lo que paulatinamente conduciría al convencimiento de que, cuando tales restricciones se cumplen, está fundamentado concluir que aquello que genera las respuestas en el organismo y lo que lo hace en la máquina pertenecen, en un sentido relevante a su explicación, a una misma clase –en

el caso de la simulación cognitiva, que la cognición y la computación son procesos homólogos, o instancias de la misma clase de proceso.

O bien *a contrario*: cabe rehusar toda concesión de valor explicativo a los modelos mecánicos de distintos procesos biológicos o psicológicos aduciendo que del hecho de que su comportamiento sea intachablemente parejo al del organismo en cuestión –o incluso, se apunta, lo perfeccione– no se sigue que sean parejos también los principios subyacentes. Ése es el fundamento de los recelos de Joseph Needham, un bioquímico e historiador de la ciencia que, unas décadas antes de que Oxford y Cambridge conocieran las máquinas de Deutsch u Oettinger, insistía en reivindicar la primacía de la investigación neurofisiológica frente a los artefactos con ínfulas epistemológicas, en tanto en cuanto "[…] the fact that these imitations of life function so well does not prove that their mechanisms are the same as those of the living brain" (Needham 1929: 153 *apud* Cordeschi 2002: 227).

Cierto exceso de perfección en el comportamiento de los autómatas, que parecía distanciarlos acaso más aun de nuestra propia propensión al error que de otros animales, y que críticos como Needham no se cansaban de señalar, pronto se iría convirtiendo en motivo de preocupación para quienes aspiraban a dar a sus autómatas valor explicativo. Acabamos de ver, por ejemplo, como Ellson (1935) se esmeraba en reproducir los errores y los tiempos de reacción observados en el laboratorio de Hull cuando los sujetos experimentales se enfrentaban a las tareas que sus modelos mecánicos trataban de abordar. El realismo psicológico era, asimismo, una de las principales inquietudes de Hugh Bradner, que en 1937 presentaría un pequeño carro capaz de aprender a recorrer un laberinto pero también de equivocarse reiteradamente al hacerlo y de –digamoslo así– despistarse como efecto de otros estímulos. También en los círculos próximos a Hull –recordemos– Ross (1938) cifraba las condiciones para poder proclamar que una máquina es capaz de aprender en que ésta mostrara idénticos patrones de relaciones entre estímulos y respuestas que el organismo cuya capacidad de aprendizaje tratamos de simular –sus errores, desde luego, no habrían de quedar fuera de la exigencia.

En el tiempo que separa a Needham de Ross[142], sin embargo, sólo se perfilan dos respuestas posibles –las suyas– a la pregunta por el alcance de la semejanza entre la imitación y lo imitado: o bien se sugiere que cualquier diferencia entre ambos –ya fuese bioquímica, anatómica, fisiológica…– invalidaría las pretensiones explicativas del modelo –puesto que, como irónicamente apuntarían Rosenblueth y Wiener (1945: 320 *apud* Cordeschi 2002: 256), "[…] the best […] model for a cat is another, or preferably the same cat"–, o bien se desdeña cualquier divergencia en lo que atañe al sustrato orgánico del comportamiento y se atiende en exclusiva a estímulos y respuestas funcionalmente caracterizados –el criterio al que luego intentaría conferir Turing (1950) carácter canónico. Aunque el imperativo de duplicar hasta el ínfimo detalle las estructuras cerebrales pareciera a ojos de muchos críticos un tanto

[142] O incluso, más tardíamente, entre Gabor (1956) y Pask (1961), *cf. supra*, cuyas valoraciones no discrepan mucho, respectivamente, de las de Needham (1929) y Ross (1938).

inmoderado –pues nos llevaría, como a Vaucanson (1738, *supra*), al afán de calcar cada apófisis de cada hueso–, y aunque dicho imperativo contraviniera –recuérdese– verdades que como había señalado el propio Ross (1938: 185, *supra*) la mecánica tomaba ya por lugar común, también para otros muchos resultaba de una lenidad desmedida aceptar como pauta de valor explicativo la mera equivalencia conductual –pues ésta, como señalara Craik (1966: 20-21) en relación con el modelo de Bradner, bien podía sustentarse en artimañas enteramente *ad hoc*, y no en la coincidencia de "principios generales", exactamente la misma objeción que Schank (1972) opondría después a *Eliza*, el famoso programa de Weizenbaum (1966) que simulaba una conversación psicoterapéutica. Pero incluso para Craik, especificar cuáles pudieran ser esos principios generales no dejaba otra salida que reclamar verosimilitud biológica. En suma, los caminos de la psicología científica –no es la primera vez que lo vemos– se agotaban en una sola encrucijada: reducción a la fisiología nerviosa o conductismo.

Tampoco el ardor dialéctico de la prolongada confrontación entre vitalistas y mecanicistas favorecía, desde luego, el tanteo de otras rutas. *Verbi gratia*: aludir, como hace Wiener (1943: 6) a que "[…] un sistema electromecánico [estaba] diseñado para usurpar funciones específicamente humanas" no resulta muy clarificador respecto a las condiciones bajo las que debiéramos admitir que tal sistema nos provee de una explicación de las funciones humanas en cuestión, o que al ponerse en funcionamiento despliega procesos homólogos de los que se dan cuando un ser humano realiza las mismas funciones. De hecho, la idea de la usurpación antes parece, por el contrario, evocar una impostura en algún sentido ilegítima. El afilado lenguaje de Wiener reescribe, en cambio, una retórica de la usurpación que ya estaba presente más de un siglo atrás, en las polémicas sobre el (falso) autómata ajedrecista que Wolfgang von Kempelen había fabricado en 1769 para divertimento de la emperatriz María Teresa de Austria, y que, exhibido por Europa y América de la mano de Johann Maelzel, el inventor del metrónomo, logró derrotar a Benjamin Franklin, a Charles Babbage, o al mismísimo Napoleón. Así, uno de los muchos críticos que se empeñaron en desenmascarar al Turco –así se conocía al presunto autómata–, Robert Willis, aseguraba en 1821 que era de todo punto imposible que una máquina pudiera hacer frente "[…] a las circunstancias siempre cambiantes del ajedrez", "[…] habilidad [que] pertenece sólo a la esfera del intelecto"; el invento de von Kempelen, por tanto, nunca podría "[…] usurpar las facultades de la mente humana" (Willis 1821: 11 *apud* Wood 2002: 70 y Guijarro y González 2010: 331). Mucho antes, hacia 1774, Pierre Jaquet-Droz, un artesano que –como había podido comprobar en la corte de Madrid, *cf. supra*– precisaba de un delicado equilibrio entre provocación y mesura para seguir vendiendo autómatas y relojes, se había deleitado, según parece, en hacer que su escribiente caligrafiara afanosamente "Cogito ergo sum". No cabe duda de que, todavía en 1943, la idea de usurpación aguijaría los ánimos de las mentalides más reacias a la mecanización de la imagen del hombre.

Después, con el tiempo, se iría abriendo paulatinamente la perspectiva de que la pregunta por las restricciones que debemos mantener en vigor para dotar de

verosimilitud y vigor explicativo a un modelo mecánico de determinado proceso psicológico puede tener distintas respuesta según el núcleo de generalizaciones del que pretendamos dar cuenta, o los objetivos de la investigación. Con inigualable concisión fijaba Pylyshyn (1979: 49) los polos de ese espectro que empezaba a dibujarse: "[…] if we apply minimal constraints we will have a Turing machine. If we apply all the constraints there may be no place to stop short of producing a human". Pero todavía en 1957, Grey Walter se hacía cargo de quienes recelaban de que el avance de la cibernética estuviera espoleando "[…] facile comparisons between living and artificial systems", y hacía depender de una vaga noción de identidad de procesos que dichas comparaciones tuvieran o no el debido fundamento:

> […] the working of a human brain has often been compared with that of a computing engine and contrariwise such instruments are often referred to as 'Electronic Brains'. In this particular case the comparison is indeed superficial and misleading; it may be based on the definition of thinking, since if this term is restricted to logical processes, then indeed the modern computing engine can perform at least as well as most human brains –but of course it remains to be proved that the procedure is the same in both. (Grey Walter 1957: 5)

Fue en el trabajo que Newell, Shaw y Simon (1958) publicarían sólo un año después donde comenzó a alumbrarse alguna manera de articular criterios bajo los que cupiera entender literalmente la analogía entre la mente y sus trasuntos mecánicos, sin dejarse llevar ni hacia la minuciosidad extravagante de un nuevo Vaucanson ni hacia la despreocupada indulgencia que había abanderado Turing. El trayecto –eso sí– nos obligaba a echar por tierra algún que otro dogma medular del conductismo, pues se trataba de recoger *verbatim* la descripción de sus propios procesos psicológicos que ofrecían los sujetos en el laboratorio mientras resolvían determinada tarea –introspección experimental, así pues, y además acerca de procesos superiores, en la estirpe de Oswald Külpe y la escuela de Würzburgo–, y después comparar esos protocolos con el registro de ejecución del programa –sus llamados "trazos"– a fin de determinar si la identidad de respuestas a idénticos estímulos iba acompañada además de identidad de procesos internos. Al mismo tiempo, la mente –o, al menos algunas de sus facultades– y los programas informáticos quedaban perfilados en un sentido laxo como elementos de la misma clase abstracta, la de *sistemas de procesamiento de información*. Siempre que la identidad de procesos internos se verificase adecuadamente, nada impediría que un programa en particular acabara por revelarse además *como miembro del mismo tipo de sistemas de procesamiento de la información que la mente* –es decir, en el sentido relevante, *como una mente*. Pronto, a partir de 1960, Putnam comenzaría a decantar del trabajo de Turing (1936, 1937) una forma algo más elegante de caracterizar ese isomorfismo funcional que Newell, Shaw y Simon habían hecho descansar sobre la comparación entre protocolos y trazos –la identidad entre tablas de máquina–, así como a descifrar las consecuencias que todo ello había de acarrear en cuanto a la credibilidad del reduccionismo, ya fuera de índole neurofisiológica o conductual.

Interludio. Autómatas y oficinistas: el cognitivismo como ideología

No estaría de más dejar apuntadas en este contexto algunas de las críticas más enérgicas que ha recibido la idea de la mente –y, por tanto, de la persona– cristalizada en el funcionalismo y en la psicología cognitiva. Una penetrante destreza para rastrear los signos e implicaciones de la recurrencia en las investigaciones cognitivistas de la noción de procedimiento mecánico ha servido a Shotter (1997), en efecto, desde una óptica radicalmente crítica con el funcionalismo, para enraizarlo no ya en la crisis fundacional de la matemática o el intenso esfuerzo de formalización y abstracción que suscitara, ni, como hace Gardner (1985), en los vertiginosos avances de la tecnología y la teoría computacional espoleados por las necesidades bélicas de la Segunda Guerra Mundial, sino en la ideología tecnocrática y burocrática descrita por Weber (1922), que subyacería también, en realidad, a lo más abominable de aquel conflicto: la meticulosa organización burocrática que hizo posible –pensable, soportable incluso para los verdugos– el Holocausto. Para Shotter, entonces, no se trata de que una reflexión sobre investigaciones que se habían desarrollado con fines bélicos en las potencias aliadas impulsara el cognitivismo, ni, desde luego, que lo hiciera una acendrada destilación de los frutos del razonamiento lógico y matemático: la cuestión es, más bien, que una misma estructura ideológica animaría tanto el detonante primero de la Guerra –el auge del nazismo–, como la paulatina deriva de la psicología, tanto experimental como filosófica, hacia posiciones cognitivistas.

La metafóra del burócrata –es cierto– se ha vuelto, al menos desde Turing (1936, 1937), una presencia acostumbrada en la teorización cognitivista; no tanto, tal vez, como la metáfora de la computadora, pero quizá más que ninguna otra. Es seguramente en los escritos de Dennett sobre la noción de homúnculo donde los burócratas comparecen más asiduamente, si bien se hallaban ya presentes, como se ha visto, en el trabajo de Turing. Como anota Shotter (1997: 323), la estrategia explicativa aconsejada por Dennett pasa por desgajar la actividad o el proceso del que pretendemos rendir cuenta:

> […] into a committee or army of intelligent homunculi with purposes, information, and strategies. Each homunculus is in turn analyzed into smaller homunculi, but, more important, into less clever homunculi. When the level is reached where the homunculi are no more than adders and substractors, by the time they need only the intelligence to pick the larger of two numbers when directed to, they have been reduced to functionaries "who can be replaced by a machine" […], [and] if the program works, then we can be certain that all homunculi have been discharged from the theory. (Dennett 1978b: 80-81)

También recuerda Shotter (1997: 323) cómo cobra forma en Broadbent (1980) la misma idea[143]:

> Imagine a man sitting in an office. On one side of him is his "in-baskets," into which people keep putting pieces of paper. […] On the other side of the man are his "out-baskets," into which he puts papers that are going to leave the office. […] If the man wants to keep information more permanently, and without blocking the use of his baskets, then he uses a filing cabinet which is placed behind him. […] [And] in addition the man has in front of him a desk on which he can put papers that he is using at the moment. (Broadbent 1980: 125)

No mucho después, Minksy (1986) proponía explícitamente las jerarquías burocráticas como modelo de la naturaleza y el funcionamiento de la mente, describiendo las operaciones de los "agentes" postulados para explicar diversos procesos psicológicos cómo "trabajo administrativo" (Minsky 1986: 34 *apud* Shotter 1997: 324).

Lo que en distintos contextos, y con distintos matices, se ha llamado solipsismo metodológico (Putnam 1975a, Fodor 1980a, *infra*), individualismo (Burge 1979, 1986, *infra*), o, por analogía con problemáticas parejas de la filosofía moral, la epistemología o la estética, internismo (Falk 1948, Frankena 1958), la idea de una semántica enteramente funcional, el carácter ahistórico de la noción cognitivista de representación[144], incluso la proclamada independencia entre los procesos psicológicos y su encarnación fisiológica, se revelarían entonces, a ojos de Shotter, como meras facetas de la ideología burocrática: "This isolation of everything from everything else typifies cognitivism" (Shotter 1997: 322). Pero es, en realidad, el compromiso naturalista del cognitivismo lo que subleva a Shotter, el hecho de que la explicación sólo se considere adecuada cuando lo intencional, lo teleológico, lo subjetivo, cuanto de humano –según su concepción de lo humano– albergue el *explandandum*, haya quedado expurgado del *explanans* –un desiderátum que es a su juicio sencillamente paradójico (Shotter 1997: 324), además de deshumanizador (Shotter 1997: 325).

Pero al igual que la mecanización de la teoría psicológica no aflora, ni mucho menos, en el cognitivismo, tampoco lo hace su burocratización: ambos parecen más bien rasgos que calan en la psicología de inspiración asociacionista desde que ésta impregna la estructura de las sociedades en que se desarrolla, y que el cognitivismo hereda de esa tradición, con la que en tantos otros aspectos litiga. Antes del apogeo de la concepción cognitivista de lo mental, quizá la expresión más nítida de ese vínculo se encuentre en los trabajos de Craik (1947, 1948) sobre el papel del operador humano en los sistemas de control automático, donde dicho operador queda

[143] Las palabras de Broadbent –cabe añadir– evocan con fuerza la escenificación del argumento de la habitación china empuñado por Searle (1980) contra la concepción más ambiciosa de la idea de inteligencia artificial, que él mismo denomina "Inteligencia Artificial Fuerte" (*supra*) –precisamente una de las objeciones al cognitivismo más conocidas.

[144] *Cf.* Cummins (1989: 84), pero también Heil (1989: 355-356), ambos *infra*.

perfilado como una cadena de tres eslabones –un dispositivo sensorial, uno computacional y uno de respuesta– que integra un elemento más del diagrama de operaciones del sistema. Desde esta óptica, tomar como metáfora a las máquinas y no a los propios computadores humanos –es decir, a los burócratas– era, retórica aparte, sólo una manera de abrir la puerta a las prometedoras perspectivas metodológicas que la simulación mecánica de la conducta, o de los procesos psicológicos, parecía ofrecer y, al mismo tiempo, poner pie en pared contra el vitalismo. En el seno del conductismo, desde luego, para poner de relieve el vínculo entre mecanización y burocratización basta contrastar –como no sin cierta malicia nos invitaban a hacer Miller, Galanter y Pribram (1960: 51-52, *supra*)– la imagen de la centralita automática que Hull empleaba sin reparos ya en 1943 con la del obsoleto telefonista que Pearson reclutara en 1892 para su explicación del funcionamiento del cerebro. Ni siquiera sería descabellado rastrear el germen de la metáfora burocrática hasta los primeros pasos de la mecanización de nuestra imagen del mundo que vino espoleada por la crisis de la física aristotélica, toda vez que en recientes análisis del procesamiento visual temprano, como es Pylyshyn (2007), podemos toparnos con palabras entresacadas de las *Ad Vitellionem Paralipomena* de Johann Kepler (1604) en las el cerebro aparece ya poblado de magistrados y cámaras administrativas:

> How the image or picture is composed by the visual spirits that reside in the retina and the [optic] nerve, and whether it is made to appear before the soul or the tribunal of the visual faculty by a spirit within the hollows of the brain, or whether the visual faculty, like a magistrate sent by the soul, goes forth from the administrative chamber of the brain into the optic nerve and the retina to meet this image, as though descending to a lower court –I leave to be disputed by [others] (Kepler 1604: 151-152 *apud* Lindberg 1976: 202 *apud* Pylyshyn 2007: 2)

Conviene subrayar, por otra parte, que la flexibilización de las jerarquías de control en las que cristaliza la metafóra burocrática ha sido una de las preocupaciones mayores de buena parte del trabajo en inteligencia artificial y simulación cognitiva desde que Newell (1962) hiciera notar su desmedida rigidez. Si bien una organización jerárquica con patrones de interacción simples y estrictamente delimitados entre unidades ejecutivas –entre rutinas y subrutinas, o, si se quiere, entre funcionarios– cuenta con evidentes ventajas –las compendia entre otros Simon (1969)–, dichas ventajas se disipan tan pronto como las unidades dejan de estar especializadas en tareas igual de simples y estrictamente delimitadas: en tales casos, como bien apuntara Newell, vale la pena disponer de canales de comunicación más dúctiles, que permitan no sólo hacer llegar al funcionario a cargo una descripción de la tarea relativamente abierta sino también inspeccionar sus progresos y tomar nota de sus propios informes al respecto[145]. La dificultad, así pues –tal como al hilo de la misma metáfora concluiría Pylyshyn (1984)– estriba precisamente en:

[145] De la misma manera, por cierto, que la idea del oficinista que subyace al trabajo de Turing parece arraigar en las primeras voces del movimiento obrero (*supra*), también la noción de jerarquía –o más bien, la protesta ante su brutalidad– pueden hallarse en las misma fuentes: que "[… los obreros] en su

> […h]ow to convert these anthropomorfically stated desiderata into mechanical form, and how to do so without swamping the system in a bureaucratic nightmare of control messages […] (Pylyshyn 1984: 80-81)

Es cierto, desde luego, que el recurso a la metafóra de la organización burocrática pervive a ese esfuerzo por atenuar sus aristas. Un ejemplo rotundo: se vale de ella el propio Newell (1973) al describir el esquema de control distribuido que conocíamos ya desde Newell y Simon (1972) como *sistema de producción*; también Pylyshyn (1984), antes de repasar sus ventajas de cara a la simulación psicológica. La presencia de la metáfora, en forma de símil, no oculta sin embargo que lo que se practica es en realidad el desmontaje de una de sus vigas maestras, la noción de jerarquía.

> A production system has two main parts –a communication area, called the workspace, and a set of condition-action pairs, called productions. If the condition side of a production is satisfied by the current contents of the workspace, then that production is said to be evoked, and the action on its action side is carried out. The workspace resembles a public bulletin board, with the productions resembling simple minded bureaucrats who do little except look at the bulletin board. Each bureaucrat, however, is looking for a certain pattern of "notices" on the board. When it finds that pattern, the "bureaucrat" performs whatever its action is […]. […T]here is, strictly speaking, no explicit control-transfer operation […]; that is, no bureaucrat ever gives any orders nor delegates authority, nor even sends messages to any other bureaucrat. All messages are broadcast […], and control is always captured by a production whose conditions happen to be satisfied by the current workspace contents. The system is completely homogeneous, distributed, and modular. (Pylyshyn 1984: 82-83)

Como no podía ser de otra manera, la visión de su propio trabajo que trasluce en los investigadores afines a los planteamientos cognitivistas es muy diferente de la bosquejada en la crítica sociológica. Valga como muestra la reflexión que cierra *Planes y Estructura de la Conducta*: enaltecer la práctica de la construcción de modelos computacionales de procesos psicológicos –la simulación cognitiva– alzándola al rango de la teorización supone renunciar al enfoque descriptivo que ha sido tradicional en ciencia en favor de un enfoque imitativo, que se ha venido considerando propio del arte. O, mejor dicho: reconocer el papel que en la trastienda de la investigación científica ha desempeñado siempre la imitación, reputarla como merece, y darle toda la primacía que permite la tecnología informática. En términos generales, los cognitivistas se han visto a sí mismos como parte de un movimiento humanizador, liberador, frente al talante severo y rígido del conductista –"the hard-nosed behaviorist" es un giro común, por ejemplo, en Baars (1986). Como se ha visto, de hecho, los lazos entre el cognitivismo y la crítica del positivismo lógico que suelen alegarse –en el marco, sobre todo, del empeño historiográfico en justificar el carácter

---

condición de soldados industriales rasos son puestos bajo la supervisión de toda una jerarquía de suboficiales y oficiales" (Marx y Engels 1848: 21-22) era una de las injurias a su dignidad que el incipiente comunismo se aprestaba a combatir.

revolucionario de la transición del conductismo al cognitivismo– son en realidad extremadamente laxos. Acaso sea, precisamente, poco más que este aire liberador lo que une los esfuerzos de los primeros cognitivistas con las nuevas orientaciones, post-popperianas, de la filosofía de la ciencia; un aire liberador que, por lo demás, se confundiría con el *Zeitgeist* de unos años marcados por la necesidad de dejar atrás el espanto de la Segunda Guerra Mundial o los rigores de la Gran Depresión que la precedió.

Tratar de sopesar con ecuanimidad la crítica sociológica del cognitivismo bosquejada por Shotter nos lleva, así pues, a constatar que las raíces de la concepción cognitiva de la mente –o de la persona– no pueden hallarse por entero en esa visión deshumanizadora que fue el substrato del nacionalsocialismo: tenemos, por una parte, que patrones de pensamiento muy semejantes se hallaban ya integrados en el conductismo, y mucho antes, y, por otra, que en la armazón conceptual y metafórica del cognitivismo juega un papel crucial la misma descripción de las tareas encomendadas a la mayor parte de los trabajadores –mecánicas, asfixiantemente jerarquizadas– que encontramos en textos medulares del comunismo, una realidad ante la cual tanto éste como el cognitivismo se presentan a sí mismos –con muy distintos tonos y en muy distintos sentidos– como movimientos emancipatorios.

Acaso en el seno de esa idea deshumanizada de lo humano el cognitivismo se conciba ilusoriamente como una liberación, cuando preserva en realidad la naturaleza opresiva de la idea de lo humano que pretende derruir –igual que, podría desearse argumentar, sucedería con la deriva totalitarista del comunismo. Nada hay, desde luego, en el presente análisis que sea óbice para esa conclusión, pero parece claro que entender cabalmente la sociología del auge de la psicología cognitiva, envuelta en tiempos tan convulsos como los marcados por las Guerras Mundiales, requeriría una reflexión mucho más detenida de la que permite este sucinto comentario.

La mudable encarnación de lo mental

Huellas tan profundas como puedan serlo las de los avances de la teoría y las tecnologías de la computación dejaría en el funcionalismo la tarea de desmantelamiento del reduccionismo fisicalista ligado a la tesis de identidad psicofísica, aunque, por ser ambos surcos coincidentes en buena parte de su derrotero, no siempre es fácil discernirlos. En la fecundísima labor de Putnam (1967a) se halla, de nuevo, el *locus classicus* en cuanto al modo en que la idea de realizabilidad múltiple –que estados mentales del mismo tipo puedan materializarse en estructuras físicas irreconciliablemente diversas entre sí– se enraiza en la crítica del fisicalismo, crítica que, como concluye Rabossi (1995: 31), "[…] está íntimamente relacionada con la prédica funcionalista". Tanto es así que es costumbre de muchos funcionalistas hacer gala del principio de realizabilidad múltiple como la contribución capital de su escuela a la comprensión de la naturaleza de la mente. La

obligación contraída por el funcionalismo de conceder el rango de estados mentales a aquellos estados computacionales que suscriban la caracterización funcional de un tipo de estado mental sería, a fin de cuentas, sólo una faceta de la idea de realizabilidad múltiple.

Entre quienes, como Bickle (1998, 2003, 2006), han adoptado una posición más crítica respecto a la tesis de realizabilidad múltiple, se tiende en cambio a diferenciar dicha tesis de las diversas conclusiones para las que ha servido de premisa: ya contrarias a la teoría de identidad psicofísica o al reduccionismo materialista en general y favorables al funcionalismo, ya, como en Putnam (1988), contrarias tanto al reduccionismo materialista como al propio funcionalismo, ya incluso, como en Kim (1992), favorables a un reduccionismo de tintes eliminacionistas. Así desligados sus términos, el argumento parece ofrecer –*divide et impera*– muchos más flancos vulnerables.

Cuando Putnam (1967a) invoca, sin darle tal nombre, a la realizabilidad múltiple de los estados mentales lo hace, en efecto, en el contexto de un explícito cotejo de las virtudes del funcionalismo y las del fisicalismo entonces al uso. Antes, Putnam ha ensayado una defensa del fisicalismo ante cuatro argumentos por los que se pretende desautorizarlo conceptualmente, *a priori*; su labor, por el contrario, apela a la verosimilitud empírica de una y otra concepción de lo mental, pero da por sentada la coherencia conceptual de ambas. No es cierto –argumenta Putnam en primer lugar– que un enunciado de identidad como "el dolor es un estado cerebral" viole regla gramatical alguna. Del hecho de que uno pueda saber que siente dolor y no saber que se encuentra en el estado cerebral correspondiente sólo se deriva que el concepto de dolor y el concepto de dicho estado cerebral son *distintos conceptos*, pero no que la propiedad de sentir dolor y la de encontrarse en dicho estado cerebral sean *propiedades distintas* ni, menos aún, que aseverar la identidad de ambas propiedades constituya esté vedado por la gramática de los enunciados de identidad. De la misma manera yerra el argumento, de acuerdo con Putnam, quien sostiene que "el dolor es un estado cerebral" no puede ser empíricamente verdadero –o sea, es necesariamente falso– porque establece una relación de identidad entre acontecimientos asociados con regiones espacio-temporales diferentes: el dolor, supongamos, está en mi brazo, pero el estado cerebral difícilmente puede estar en mi brazo. Tal vez resulte obvio que lo que parece subyacer a estos dos intentos fallidos de desarbolar el fisicalismo antes de que llegue a zarpar es un empleo ilegítimo de la ley de Leibniz. En ambos casos, la indiscernibilidad de los idénticos, la sustituibilidad de expresiones correferenciales *salva veritate*, se toma como premisa para dar por absurda, en uno u otro sentido, la tesis de identidad. Pero se obvia el hecho de que en ambos casos los enunciados en los que se basa la impugnación de la tesis de identidad psicofísica forman contextos opacos, en los que el término psicológico y el neurológico aparecen en *oratio obliqua* y en los que, por tanto, el quebrantamiento de la ley de Leibniz es esperable. O al menos, cabe argumentar que así sucede: es cuestionable que "saber", cuando se usa para afirmar que uno sabe que siente dolor, o "estar", cuando se usa

para afirmar que el dolor de uno está en el brazo de uno, se comporten, en cuanto atañe a la ley de Leibniz, de forma diferente que "creer" en la afirmación de que uno cree sentir dolor o cree que el foco del dolor se encuentra en el brazo; en consecuencia, es cuestionable que, a diferencia de "creer", introduzcan contextos referencialmente transparentes. Así vista, la táctica de Putnam para combatir la primera crítica que afronta –a saber, la táctica de diferenciar el concepto de dolor de la propiedad de dolor, y poner a salvo a esta última– aparecería como una versión de la estrategia general esbozada: lo que Putnam logra imponer en la discusión al distinguir el concepto de dolor de la propiedad de padecer dolor es el componente intensional del concepto "concepto de dolor", que abre la puerta a los mencionados contextos opacos, dejando a buen recaudo la extensionalidad del concepto "propiedad de dolor", que queda impoluto para su empleo en enunciados de identidad de forma plenamente acorde con la ley de Leibniz.

Por motivos semejantes, no es cierto tampoco que enunciados del estilo de "el dolor es un estado cerebral" sean ininteligibles o carentes de sentido, no más ininteligibles o carentes de sentido, en todo caso, que cualquier otro enunciado de identidad que pueda germinar en la investigación empírica, como "el agua es H_2O". Por último, Putnam parece concidir con Smart (1959) en que resulta un tanto desesperado argumentar que cualquier evidencia empírica que pudiera aportarse en defensa de la tesis de identidad psicofísica quedaría igualmente suscrita por la tesis de que los estados psicológicos, lejos de *ser* estados cerebrales, presentan una correlación perfecta con ellos. Como sagazmente plantea Putnam la cuestión, la razón de que ante tal paridad explicativa nos inclinemos por la hipótesis de identidad, más sencilla, no es un capricho ockhamiano, sino el hecho de que esa hipótesis expurga nuestro proyecto de investigación de preguntas estériles como la de cuál es entonces la verdadera naturaleza de lo mental, o cuáles son las causas de la feliz correspondencia que guarda con lo cerebral –desprovee a esas preguntas de "significación empírica" (Putnam 1967a: 225)–, y nos evita así volver a recorrer caminos que ya fatigaron Malebranche, Spinoza o Leibniz.

Una vez que estos cuatro intentos de arruinar de antemano el proyecto fisicalista han quedado contrarrestados, pues, Putnam se dispone a examinar funcionalismo y fisicalismo en pie de igualdad conceptual y en razón de su vigor empírico:

> Indeed, my strategy will be to argue that pain is *not* a brain state, but not on *a priori* grounds, but on the grounds that another hypothesis is more plausible. […] I propose the hypothesis that pain, or the state of being in pain, is a functional state of a whole organism. (Putnam 1967a: 226)

La tesis de realizabilidad múltiple aparece como la clave de ese escrutinio, el plomo que tercia la balanza del lado del funcionalismo. La crítica del fisicalismo y la "prédica funcionalista" se dan en Putnam sin solución de continuidad: a la premisa de que el teórico fisicalista debe dar con un estado cerebral tal que, *para todo*

organismo, el organismo siente dolor si y solo si se encuentra en dicho estado cerebral (*cf*. Liz 1995: 223, *supra*), y replicar tamaño descubrimiento para todas las criaturas físicamente posibles y para todos los demás estados psicológicos –un presupuesto acerca de la naturaleza de los procesos de reducción interteórica que los críticos de Putnam cuestionarán– se une la de que cumplir ese requisito es a duras penas posible, a la luz de unas sencillas consideraciones evolutivas; de inmediato, se plantea que la exigencia en cuestión resulta incomparablemente más asequible cuando de lo que se trata es de fijar un estado funcional tal que, para todo organismo, el organismo siente dolor si y solo si se encuentra en dicho estado estado funcional, y replicar tamaño descubrimiento, etc., debido a que las consideraciones evolutivas pesan ahora en favor de la empresa, y a que a ellas se añaden otras relativas a nuestros propios criterios de identificación de estados psicológicos. Dicho de otro modo: cuando el defensor de la identidad psicofísica afronta la labor de aislar un tipo de estado cerebral que se registre en todo organismo o sistema que sienta dolor y únicamente en esos organismos o sistemas, tiene en su contra ciertos hechos empíricos poco controvertidos acerca de la evolución de las especies y las múltiples ramas en que ésta se despliega simultáneamente:

> Even though octopus and mammal are examples of parallel (rather than sequential) evolution, for example, virtually identical structures (physically speaking) have evolved in the eye of the octopus and in the eye of the mammal, notwithstanding the fact that this organ has evolved from different kinds of cells in the two cases. Thus it is at least possible that parallel evolution, all over the universe, might *always* lead to *one and the same* physical "correlate" of pain. But this is certainly an ambitious hypothesis. (Putnam 1967a: 228)

En cambio, el funcionalista gana el barlovento cuando aborda esa misma tarea –aislar un tipo de estado funcional presente en todo organismo o sistema que sienta dolor y únicamente en ellos–, ya que la propiedad cuyo equivalente funcional trata de hallar –el dolor, la sed, o el estado mental que sea– es una propiedad cuya presencia en un organismo o sistema *identificamos* atendiendo precisamente a la conducta del organismo o sistema en cuestión –a su conducta, cabría añadir, ante determinados estímulos. Así, arguye Putnam (1967a: 229), ni siquiera consideraríamos que un animal está sediento si la conducta subsiguiente no resultara encaminada a la ingesta de líquido, y dicha ingesta no elicitara un estado de saciación; *mutatis mutandis*, lo mismo sucede en el caso del dolor, o en el de cualquier estado psicológico. Que los tipos de estados psicológicos casaran con tipos de estados funcionales, así pues, no sería lo insólito –como en la identidad psicofísica–, sino que, por regla general, sería lo esperable:

> […] it is a truism that similarities in the behavior of two systems are at least a reason to suspect similarities in the functional organization of the two systems, and a much *weaker* reason to suspect similarities in the actual physical details. Moreover, we expect the various psychological states –at least, the basic ones, such as hunger, thirst, aggression,

> etc.– to have more or less similar "transition probabilities" (within wide and ill-defined limits, to be sure) with each other and with behavior in the case of different species, because this is an artifact of the way in which we identify these states. (Putnam 1967a: 228-229)

La capacidad que se ha atribuido al funcionalismo de aunar lo más ventajoso del conductismo lógico y de la tesis de identidad psicofísica, evitando a la par lo menos prometedor de ambos enfoques, puede ahora volver a ser glosada, desde esta óptica, en palabras de Fodor:

> Viewed from one perspective, [...] [functionalism] offered behaviourism[146] without reductionism; viewed from another, it offered physicalism without parochialism. The idea that mental particulars are physical, the idea that mental kinds are relational, the idea that mental processes are causal, and the idea that there could, at least in logical principle, be nonbiological bearers of mental properties were all harmonized. And the seriousness of the psychologist's undertaking was vindicated by providing for a Realistic interpretation of his theoretical constructs. (Fodor 1981a: 10-11)

Es crucial advertir la asimetría entre el papel desempeñado por la tesis de realizabilidad múltiple en el argumento contra el fisicalismo y el que juega en el argumento en favor del funcionalismo. El propio Putnam presenta la tesis de realizabilidad múltiple de los estados psicológicos como una hipótesis empírica: una hipótesis con muchos visos de ser verdadera, cuya negación es extremadamente ambiciosa y –cree Putnam– a todas luces falsa, pero una hipótesis al fin y al cabo. A su juicio, encontrar un único caso en que diéramos por verdadera la tesis de realizabilidad múltiple conllevaría de inmediato la falsedad del fisicalismo, entendido bajo los parámetros de la teoría de identidad psicofísica con alcance de tipos, y resulta "[…] abrumadoramente probable que podamos hacer tal cosa":

> […] if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say "hungry"), but whose physical-chemical "correlate" is different in the two cases, the brain-state theory has collapsed. (Putnam 1967a: 228)

El razonamiento que Putnam aduce en defensa del funcionalismo, por el contrario, reposa sólo indirectamente sobre la aparente verosimilitud empírica de la tesis de realizabilidad múltiple[147]. Una vez que dicha verosimilitud ha servido para dar por establecida la falsedad del fisicalismo, la plausibilidad de la alternativa funcionalista parece depender de un argumento de orden aparentemente conceptual, o cuasi-conceptual, que atañe al hecho de que los criterios de identificación que usamos para atribuir estados psicológicos a otros organismos tienen que ver con su conducta. Ese

[146] Más exactamente: "[…] a relational construal of mental properties", *cf*. Fodor (1981a: 10, *supra*).

[147] También Bickle (2006: §1.3) señala el carácter indirecto del respaldo que la realizabilidad múltiple presta al funcionalismo, si bien cifra la diferencia entre el ataque de Putnam al fisicalismo de tipos y su defensa del funcionalismo en el carácter deductivo (y válido) del argumento que subyace al primero, por oposición a la naturaleza abiertamente no-deductiva del argumento en favor del funcionalismo.

hecho, naturalmente, es un hecho empírico: aunque tratándose de explicar lo que nos lleva a adjudicar sensaciones, sentimientos o pensamientos a otros seguramente estuviéramos condenados al fracaso de pretender hacerlo mediante la sola apelación a la aprehensión introspectiva de las cualidades fenomenológicas de nuestros propios estados mentales, podría darse el caso de que de hecho identificáramos los estados psicológicos de otros organismos en virtud de una finísima percepción de ciertos cambios fisiológicos en el otro, o de una capacidad innata de escrutinar otras mentes con la misma o mayor facilidad que la propia, o tal vez de un oráculo… Pero esas hipótesis son tan extravagantes que no parece particularmente aventurado asumir que el recurso a indicios conductuales para atribuir estados mentales a otros organismos forma parte de la naturaleza de nuestro sistema conceptual, y argumentar desde esa base –concedamos– cuasi-conceptual[148]. Con palabras de Putnam (1967a: 229, *supra*), en efecto, lo que pesa en favor del funcionalismo es "[…] un artefacto del modo en que identificamos" estados mentales –no, pues, evidencia que sustente una hipótesis empírica.

Desde esta óptica, sin embargo, podría parecer que Putnam ha hablado en defensa del conductismo aunque pretendiera hacerlo en defensa del funcionalismo. Es al fin y al cabo a la conducta de los organismos a lo que apelamos para atribuirles estados mentales, así que, según las reflexiones del propio Putnam, eso debería inclinarnos hacia un análisis disposicional de dichos estados mentales, *à la* Ryle, más que hacia un análisis en términos de su organización funcional. Deshacer esa interpretación de su argumento es el siguiente empeño de Putnam, y la distinción entre conceptos y propiedades, de nuevo, lo cardinal de sus pertrechos. Que identifiquemos el dolor, o cualquier otro estado mental, atendiendo a las conductas a las que nos predispone dice seguramente mucho sobre *nuestro concepto* de dolor, o del estado mental en cuestión, pero apenas nada sobre la naturaleza de la *propiedad* de sentir dolor, o de atravesar el estado mental del que se trate. El elemento de intensionalidad –de opacidad referencial– introducido por medio de la noción de "concepto" es ahora más patente si cabe:

> […] it is necessary to our hypothesis that the marks that are taken as behavioral indications of pain should be explained by the fact that the organism is in a functional state of the appropriate kind, but not that speakers should *know* that this is so. (Putnam 1967a: 229)

Ahora bien, ¿no podría el fisicalista remedar esta táctica de Putnam ante el conductismo para minar la defensa del funcionalismo del propio Putnam? Es decir, ¿no es poco menos que inevitable que el fisicalista, como haría Lewis (1969: 232, *infra*), argumente que nuestras prácticas de identificación de estados mentales no conciernen a la naturaleza de la propiedad de atravesar dichos estados, sino a

[148] Las consideraciones de Quine (1953) o del propio Putnam (1962) acerca de lo artificioso de la dicotomía analítico–sintético están, como tal vez resulte obvio, en el trasfondo de esta difuminación de las lindes entre un argumento conceptual y uno de naturaleza empírica.

nuestro concepto de ellos –que argumente que Putnam incurre en la misma confusión entre conceptos y propiedades que denuncia? Dicho de otro modo, parece que la derrota de Putnam lo escora bien hacia un argumento contra el fisicalismo que lo deja inerme ante el conductismo, bien hacia uno contra el conductismo que lo deja, de nuevo, inerme ante el fisicalismo[149].

En la medida en que descansan sobre la verosimilitud empírica de la tesis de realizabilidad múltiple, sin embargo, las objeciones al fisicalismo enarboladas por Putnam quedan intactas. Al otro costado, el rechazo del análisis disposicional por parte de Putnam (1967a) no estriba únicamente en la distinción entre conceptos y propiedades, sino que se despliega como un reensayo de la implacable crítica del conductismo que había quedado ya articulada en Putnam (1963a) –a saber: la totalidad de las disposiciones conductuales correspondientes a un determinado estado mental puede, concebiblemente, presentarse en ausencia de éste, así como éste en ausencia de aquéllas. Además, a las disquisiciones modales ensayadas en el ensayo original –super-espartanos, super-super-espartanos y super-actores–, Putnam (1967a) prefiere sobreescribir otras de aire más empírico, relativas a las secuelas conductuales y psicológicas de diversas, hipotéticas, neurectomías cuyos efectos serían semejantes a los de la prodigiosa fuerza de voluntad de los super-espartanos, aunque dramáticamente menos específicos:

> For example, two animals with all motor nerves cut will have the same actual and potential "behavior" (viz., none to speak of); but if one has cut pain fibers and the other has uncut pain fibers, then one will feel pain and the other won't. (Putnam 1967a: 230)

Así pues, la fuerza de su reprobación del conductismo lógico –sea cual sea– no se ve menguada por la travesía del angosto paso al que Putnam se abocara al emplear contra el conductista consideraciones que el fisicalista puede hacer virar en contra de la propuesta funcionalista con la que el propio Putnam pretende desbancarlo.

En suma, no sería aventurado concluir de forma tentativa que, aun dejando indemnes sus críticas del fisicalismo y el conductismo, esta constatación dañaría gravemente, si no se le pone remedio, los argumentos de Putnam en pro del funcionalismo, puesto que limita su alcance al de la elucidación del concepto de dolor –sed, hambre…– a la vez que los desprovee de toda autoridad sobre la explicación de la naturaleza de las propiedades relevantes. Pero se ha dado por bueno que el impulso que Putnam (1967a) pretendía dar al funcionalismo era de índole conceptual, o acaso cuasi-conceptual en el sentido vagamente esbozado: no es raro, pues, que eso nos haya llevado a concluir que sus frutos pertenecen al ámbito del esclarecimiento conceptual.

---

[149] No es extraño, a la vista de todo esto, que Lycan (1987: 8) llegue a concebir las propuestas de Putnam (1967a) como una *vuelta hacia el conductismo*, encaminada a corregir el exagerado énfasis en la identidad psicofísica que había caracterizado la reacción de los fisicalistas ante las dificultades que atenazaban al propio conductismo.

Hay, sin embargo, más que decir contra ese modo de interpretar los argumentos de Putnam. No en vano, él mismo abre su trabajo con una defensa de la tesis de identidad psicofísica ante argumentos que pretenden desacreditarla *a priori*, describiendo su propia estrategia como la de "[…] argumentar que el dolor *no* es un estado cerebral, no sobre una base *a priori*, sino sobre la base de que otra hipótesis es más plausible" (Putnam 1967a: 226); del mismo modo, apenas unas páginas después, cierra su contraposición de fisicalismo y funcionalismo apelando a futuros desarrollos empíricos de la psicología razonablemente esperables –en particular, a la formulación de leyes psicológicas supraespecíficas– como un horizonte propicio para las tesis funcionalistas. Para el fisicalista, en cambio, escudriñar un horizonte parejo sería, a juicio de Putnam, poco más que haber vislumbrado un espejismo:

> […] if the program of finding psychological laws that are not species-specific –i.e., of finding a normal form for psychological theories of different species– ever succeeds, then it will bring in its wake a delineation of the kind of functional organization that is necessary and sufficient for a given psychological state, as well as a precise definition of the notion "psychological state." In contrast, the brain-state theorist has to hope for the eventual development of neurophysiological laws that are species-independent, which seems much less reasonable than the hope that psychological laws (of a sufficiently general kind) may be species-specific, or, still weaker, that a species-independent *form* can be found in which psychological laws can be written. (Putnam 1967a: 229)

Ahora bien, que tales resultados empíricos pudieran alentar las conclusiones de Putnam se compadece mal con la idea de que estas sean enteramente conceptuales, o cuasi-conceptuales, en su origen y naturaleza. Desde esta perspectiva, resulta exagerado –y contraproducente– el énfasis del propio Putnam (1967a: 229, *supra*) en que la ventaja del funcionalismo sobre el fisicalismo radica en "[…] un artefacto del modo en que identificamos" estados mentales. Aunque tal artefacto parece existir, aunque tal vez verdaderamente descalabre las ambiciones fisicalistas, el caso es que abriga acaso mejor al proyecto conductista que al funcionalista, por un lado, y, por otro, que el funcionalismo de Putnam no depende crucialmente de él, en la medida en que –como después el de Fodor o el de Block– aspira a sustentarse sobre el desarrollo científico de la psicología y no –como el de Lewis– sobre el análisis del discurso ordinario. De hecho, precisamente Block y Fodor (1972a: 46, *infra*), se refieren a los argumentos de Putnam que, a juicio de ambos, habrían liquidado la credibilidad del fisicalismo como "consideraciones empíricas"; el propio Putnam, en el penúltimo párrafo de "La naturaleza de los estados mentales" recapitula sus razonamientos bajo el rubro de "[…] razones empíricas para decir que sufrir dolor es un estado funcional, y no un estado cerebral o una disposición conductual" (Putnam 1967a: 230). Lo que ha presentado, sin embargo, no es eso, sino razones empíricas para decir que sufrir dolor no es un estado cerebral, razones conceptuales para decir que no es una disposición conductual, y razones igualmente conceptuales, o cuasi-conceptuales, para decir que es un estado funcional.

De todos modos, como acertadamente hacía notar Block (1978), la razón por la que la tesis de realizabilidad múltiple constituye un motivo para desconfiar del fisicalismo *no es* que los estados mentales puedan darse en sistemas cuya composición y estructura física sea sumamente diferente. Las propiedades físicas, en general, pueden darse en objetos tan físicamente dispares como queramos: también de un cerebro y de una computadora –o de cualquier otra cosa– se predica, por ejemplo, que se encuentran a una determinada temperatura. Si ello no convierte a la temperatura en un fenómeno refractario al análisis fisicalista es porque *damos por hecho* que hay una determinada propiedad física –*grosso modo*, para un gas, la energía cinética media de las moléculas que los componen– que esos objetos dispares comparten, y que constituye el fundamento de nuestra atribución a todos ellos de idéntica temperatura. El problema en el ámbito de los estados mentales es de otro orden: no es sólo que desconozcamos esas propiedades físicas fundamentales que habrían de darse únicamente allí donde se diera un estado mental de cierto tipo, sino que resulta *prima facie* extremadamente improbable que tales propiedades pudieran existir –o, tal como prefiere plantearlo Block, que ni siquiera nos es dado concebir de qué propiedades habría de tratarse:

> […T]he functionalist argument against physicalism is that it is difficult to see how there *could be* a nontrivial first-order […] physical property in common to all and only the possible physical realizations of a given Turing-machine state. […] At the very least, the onus is on those who think such physical properties are conceivable to show us how to conceive of one. (Block 1978: 65)

El giro crucial en la argumentación de Block parece eludir a Bickle (1998, 2006) cuando éste intenta acomodar la realizabilidad variable radical –es decir, aquellos casos en los que se encontrasen en el mismo individuo a lo largo del tiempo estados mentales del mismo tipo cuyas distintas materializaciones resultaran irreconciliablemente diversas– como un rasgo habitual en cualquier proceso de reducción teórica, que quedaría así desprovisto de aristas problemáticas. Es precisamente en la reducción de la temperatura de un gas a su energía cinética media donde Bickle cree dar con un ejemplar domesticado de realizabilidad variable en su variedad más feroz:

> For any token aggregate of gas molecules, there is an indefinite number of realizations of a given temperature—a given *mean* molecular kinetic energy. Microphysically, the most fine-grained theoretical specification of a gas is its microcanonical ensemble, in which the momentum and location (and thus the kinetic energy) of each molecule are specified. Indefinitely many distinct microcanonical ensembles of a token volume of gas molecules can yield the same *mean* molecular kinetic energy. Thus at the lowest level of microphysical description, a given temperature is vastly multiply realizable in the same token system over times. Nevertheless, the case of temperature is a textbook case of reduction. So this type of multiple realizability is not by itself a barrier to reducibility. (Bickle 2006: §2.6)

Pero la piedra angular del razonamiento de Block es precisamente que el terreno de las ciencias del cerebro es yermo en conceptos equiparables no ya al de ensamblaje microcanónico, sino al de energía cinética media, y que ni siquiera tenemos una vaga noción de cuál podría ser la fisonomía de esos conceptos. Que los mecanismos moleculares de la facilitación y la consolidación de la memoria a largo plazo, imbricados en la cascada de señales bioquímicas que gobierna la actividad plástica de las sinapsis en respuesta a la experiencia, puedan resultar universales –el estudio de caso que Bickle (2003, 2006) toma como baluarte– constituye sin lugar a dudas un avance colosal en nuestro conocimiento del sistema nervioso. Sin embargo, el paso desde ese hallazgo a la conclusión de que "[…] we should expect that the molecular mechanisms for any causally efficacious cognitive kind be 'universally conserved'" (Bickle 2006: §2.7) no es, ni mucho menos, tan expedito como Bickle lo describe, ni siquiera apelando al principio general de evolución molecular según el cual ésta tiende a ser más lenta cuanto mayor es la relevancia funcional del dominio al que afecta, y particularmente lenta cuando se trata de proteínas de mantenimiento, implicadas en el metabolismo celular de muchos tejidos orgánicos diferentes. Lo más lejos que esta línea de razonamiento puede llevarnos es, precisamente de la mano de Block (1997, *infra*), a la convicción de que, dado que a esas fuerzas restrictivas se suma en niveles más elevados de descripción el efecto de otras igualmente centrípetas, como la selección natural o los procesos de aprendizaje, cabe esperar mayor uniformidad cuanto más alto sea el nivel descriptivo adoptado. Así que estaríamos al fin y al cabo ante una reivindicación –matizada, pero una reivindicación– de las tesis funcionalistas.

Por lo demás, es precisamente debido, al menos en parte, a las diferentes fuerzas que operan en diferentes ámbitos de la realidad por lo que es de esperar que un vocabulario teórico de orden superior nos permita generalizaciones impracticables con los recursos explicativos que resultan idóneos en disciplinas básicas –lo cual no significa necesariamente que nos proporcione relaciones entre clases más abstractas o más groseras, sino, a veces, como se verá en relación con las protestas de Bechtel y Mundale (1999: 105, *infra*), primorosamente más finas. Lograr, como anhela Bickle (2003, 2006: §2.7), una "reducción despiadada" de la consolidación mnemónica a la vía de señales moleculares en la que intervienen el adenosín monofosfato cíclico, la proteína quinasa A, y las proteínas reguladoras $CREB_1$ y $CREB_2$, exigiría remedar, mediante el vocabulario propio de la bioquímica, sin merma de significado alguna, cualquier generalización que sobre la pervivencia de los recuerdos en la memoria pudiera ofrecernos la psicología. Tal vez eso no sea imposible *en principio*, pero conviene no perder de vista la envergadura de la tarea.

Más tajante incluso se muestra Rabossi (1995), que reemplaza la dificultad de concebir las propiedades que Bickle rastrea por la imposibilidad de que existan, la apelación a la carga de la prueba por la retórica de la demostración. Al describir el argumento de la realizabilidad variable como clave del "descrédito" del fisicalismo y

la consiguiente "entronización del funcionalismo", Rabossi deja claro que, a su entender, dicho argumento sencillamente

> […] niega que sea posible identificar los estados psicológicos con estados neurológicos, porque su relación no es de de uno a uno sino de uno a muchos. El punto es que no existe, ni puede existir, una clase natural física única que pueda correlacionarse con cada clase natural genérica de la psicología […]. Las propiedades psicológicas se realizan (instancian, implementan, ejemplifican) en bases físicas heterogéneas. En consecuencia, no pueden ser identificadas (ni son identificables) con una de ella ni con la disyunción de todas ellas. (Rabossi 1995: 36)

El giro categórico dibujado por Rabossi responde a la incisiva distinción que perfila entre una versión empírica de la tesis de realizabilidad múltiple y una versión conceptual, cuyas vertientes venimos ya tanteando. De acuerdo con esta última, en efecto, la realizabilidad múltiple no es sino una consecuencia lógica de tipo de propiedades que identificamos como propiedades psicológicas y del modo en que lo hacemos:

> Las propiedades psicológicas, concebidas como propiedades causal/funcionales, son propiedades de segundo orden (es decir, son propiedades de las propiedades de los estados mentales), y la especificación de tal tipo de propiedades no impone restricciones en cuanto a su implementación física. (Rabossi 1995: 37)

En la interpretación conceptual del argumento, así pues, la distinción entre las objciones antifisicalistas de Putnam, de corte empírico, y su defensa del funcionalismo, de corte conceptual, quedaría colapsada. Sería en realidad el mismo análisis conceptual que pesa en favor del funcionalismo el que dotaría de credibilidad a la tesis de realizabilidad múltiple (mejor dicho, recordemos: el mismo análisis cuasi-conceptual que parece pesar más bien en favor del conductismo, requiebro que Putnam intenta evitar mediante un argumento que el fisicalista, en un nuevo giro contra Putnam, puede hacer suyo). La tesis de realizabilidad múltiple, entonces, no sería ya la afirmación presuntamente empírica de que resulta en extremo inverosímil que todos y cada uno de los casos que clasificamos como pertenecientes al mismo tipo de estado mental compartan un conjunto distintivo de características neurofisiológicas, o físicas, sino más bien la afirmación conceptual de que nuestra idea de los estados mentales es tal que ese grado de vaguedad en lo que atañe a su encarnación le resulta inherente.

Funcionalismos: cartografía teórica

Las generalizaciones respecto a las relaciones con estímulos, respuestas y otros estados internos que sirven para definir un estado funcional pueden provenir de una teoría psicológica desarrollada en el seno de la ciencia, o del conocimiento psicológico que forma parte del sentido común. Es habitual distinguir, en función de

esa diferencia en el linaje de las relaciones funcionales reivindicadas en la teoría, entre un funcionalismo del sentido común, analítico o conceptual, y un funcionalismo aplicado a la psicología científica, que Block (1978, 1980a) bautizó como "psicofuncionalismo". El funcionalismo analítico –al que Block (1978: 67) se refiere asimismo como "funcionalismo *a priori*" anidaría en los trabajos de Smart, Lewis y Armstrong –también de Shoemaker–, mientras que distintas versiones de un funcionalismo "empírico" habrían sido defendidas con mayor o menor convencimiento por Putnam, Fodor, Harman y el propio Block[150]. Así pues, según el funcionalismo empírico "mental" y "psicológico" son coextensivos, y las propiedades que la psicología científica asigne a los fenómenos psicológicos son de hecho las propiedades de lo mental; según el funcionalismo empírico, además, la psicología científica convendrá en una taxonomía funcional de tales fenómenos. Para el funcionalismo analítico, por el contrario, las propiedades de lo mental no vendrán perfiladas por el fatigoso avance de la ciencia, sino por el *corpus* de conocimientos al que intuitivamente recurrimos día a día cuando tratamos de entender el comportamiento de los demás, y a menudo, acaso, también el nuestro; según el funcionalismo analítico, además, esta psicología natural está construida sobre cimientos funcionales.

Una manera sistemática de abordar la controversia entre la vertiente empírica y la vertiente analítica del pensamiento funcionalista es examinar a fondo el uso que en cada una se hace de los conceptos de rol funcional (o causal), ocupante de ese rol,

[150] De muy otra índole es la diferencia entre la concepción del funcionalismo como una tesis de identidad de estados funcionales y como una tesis de identidad de especificaciones funcionales, que –como ha de examinarse más pausadamente *infra*– sirve a Block (1980a: 38-39) para tratar de elucidar el enfrentamiento, en el seno del funcionalismo, entre la idea de que éste muestra la verdad del fisicalismo y la de que, como opina el propio Block, muestra su falsedad. Por regla general, sin embargo, la interpretación del funcionalismo como una tesis de identidad de especificaciones funcionales, con los compromisos ontológicos fisicalistas que de ello derivan, ha ido de la mano de la idea de que las generalizaciones que dieran contenido a la caracterización funcional de nuestros estados y procesos psicológicos deberían provenir de la psicología cotidiana o el sentido común, para cuyos conceptos mentalistas ofrecerían así un análisis filosófico. En cambio, la interpretación del funcionalismo como una tesis de identidad de estados funcionales, ontológicamente neutral, ha tratado de colmar sus caracterizaciones funcionales con generalizaciones avaladas por el trabajo científico, con o sin la esperanza de que esta minuta acabara por iluminar también la naturaleza de los conceptos mentalistas que saturan nuestra vida diaria. Sea como sea, parece claro que adoptar un enfoque analítico o uno empírico de la caracterización funcionalista de los estados mentales e interpretar tales caracterizaciones desde una posición celosamente fisicalista o desde un comedido equilibrio ontológico son, al menos en principio, decisiones independientes.

Una posición más radical a este respecto ha sido defendida por Martínez-Freire (2001: 94), quien argumenta que considerar las tesis de Lewis y Armstrong como una variedad de funcionalismo es simple y llanamente un error. El propio Lewis, como recuerda Hierro-Pescador (2002: 122), ha expresado reiteradamente sus dudas respecto de su condición de funcionalista, motivadas en buena medida por su convicción de que los nombres de estados mentales designan los estados físicos que ocupan un determinado papel causal en una clase particular de organismos o sistemas. Para Hierro-Pescador, sin embargo, no hay razones de peso para excluir a Lewis de la nómina funcionalista.

propiedad funcional, propiedad de orden superior, y otros afines. En una primera aproximación, verter sobre los conceptos que empleamos para mencionar estados mentales la nítida distinción entre el rol causal mediante el que delimitamos la extensión de un concepto y los distintos ocupantes de ese rol causal –distintos, se entiende, en cualquier conjunto de propiedades que no impida que desempeñen el rol en cuestión– tiene como resultado el núcleo de la concepción de lo mental del funcionalismo:

> […E]l concepto de un estado mental [del] tipo *x* […] es el concepto de algo que de un modo característico causa determinados efectos y es a su vez efecto de ciertas características. Los efectos de [un estado mental del tipo] *x* son patrones de conducta de la persona que tiene [el estado mental del tipo] *x*. Las causas de [un estado mental del tipo] *x* son eventos en el entorno de esa persona. Una descripción del rol causal de *x* incluye la descripción de sus causas y efectos típicos, así como las conexiones causales posibles de *x* con otros estados mentales. Adviértase que lo que se especifica es el rol causal que *x tiene* en su condición de mediador interno entre las causas del entorno y los efectos conductuales. No se afirma que *x sea* ese rol causal. (Rabossi 1995: 34)

Entre paréntesis: la crucial diferencia entre que algo *tenga* un determinado rol causal y *ser* un determinado rol causal que algo tiene –apunta Rabossi– marcaría la distancia entre el funcionalismo analítico y el conductismo; al lado –se da por obvio– de la mención a "las conexiones causales posibles de *x* con otros estados mentales". Ambos asuntos están más entretejidos de lo que parece a primera vista. Algo que *tiene* un determinado rol causal bien puede contar, entre las relaciones causales que definen dicho rol, con lazos con cualquier otro poseedor de un rol causal, pero en cambio es lógicamente insostenible que el rol causal de algo tenga relaciones causales con el rol causal de otra cosa. Si un dispositivo mecánico tiene el rol causal de *carburador* en un motor de combustión interna, bien pudiera ser que mantuviera relaciones causales con otros dispositivos mecánicos, como los cilindros, que poseyeran otros roles causales –de hecho, difícilmente podría cumplir su cometido de otro modo. Pero el rol causal en cuestión –*grosso modo*: mezclar aire y carburante generando el combustible que se transporta a los cilindros– no puede establecer relaciones causales con otros roles –como el de producir la ignición del combustible en cada cilindro, que desempeñan las bujías–, simplemente porque un rol causal no es una cosa que entable relaciones causales, sino un conjunto de relaciones causales que es característico de una cosa. Por trasladar la metáfora automovilística a un terreno más sencillo: es obvio que uno puede hablar con una persona que desempeña la tarea de ordenar el tráfico, pero no puede hablar, se ponga como se ponga, con dicha tarea. Este aparente bizantinismo da cuenta de la convicción de Lewis (1966) de que el funcionalismo, como la tesis de identidad psicofísica pero a diferencia del conductismo, lograba investir de plena realidad a los estados mentales, sus causas y sus efectos; también del dictamen de Fodor (1981a) respecto a que el funcionalismo habría aprendido del fisicalismo la lección de la "autonomía ontológica de los particulares mentales y […] el carácter causal de las interacciones mente-cuerpo"

(Fodor 1981a: 9, *supra*), y del conductismo la lección de la naturaleza relacional de las propiedades psicológicas.

La distinción entre un rol causal y el inquilino que lo profesa sirve también para articular, como hiciera Armstrong (1968), el proyecto de investigación propio del funcionalismo analítico, en el cual una primera fase, de orden conceptual y apriorístico, que "[…] ofrece análisis detallados de los diferentes conceptos mentales tipo, explicitando así su significación", precede a una segunda, empírica, que "[…] ubica y describe los estados físico-químicos tipo del cerebro que ocupan los roles causales atribuidos a los conceptos mentales" (Rabossi 1995: 34). Por otra parte, se hace patente así la cercanía ya apuntada entre el funcionalismo analítico y las formulaciones pioneras de la tesis de identidad psicofísica: según el funcionalismo analítico, el provecho que lenta y fatigosamente iremos obteniendo de la investigación vendrá dado por la reiterada confluencia de la definición analítica del rol causal de un tipo de estado mental *M* y la delimitación empírica del rol causal de un tipo de estado cerebral *N*, confluencia que nos autorizará razonablemente a concluir que *M* = *N* y, a la postre, que la mente y el cerebro son una y la misma cosa. En la acertada síntesis de Rabossi (1995), quien de hecho registra el funcionalismo analítico como una variante de la tesis de identidad psicofísica y prefiere denominarlo "teoría de identidad de rol causal":

> Una teoría, la psicología de sentido común, permite la introducción de términos caracterizados por su rol causal. Otra teoría, la neurofisiología, en conjunción con la primera, implica las identidades psicofísicas. El significado de los términos pertinentes y la neurofisiología conducen, necesariamente, a las identidades psicofísicas. (Rabossi 1995: 35)

En el marco, dicho sea de paso, de una severa crítica del funcionalismo empírico, que incurriría a sus ojos en las mismas faltas que desacreditaron al conductismo lógico y al fisicalismo de tipos, Block (1978: 78-80) desplegó tres vigorosas objeciones contra el funcionalismo analítico. Además de compartir los desarreglos que –si Block está en lo cierto– dañarían el esquema explicativo del funcionalismo empírico, la versión *a priori* de la teoría está infectada de males mayores. Así, Lewis está firmemente comprometido con la idea de que las identificaciones funcionales se fundamentan en el análisis del significado del vocabulario de la psicología cotidiana; esto le fuerza a asumir que las generalizaciones de las que se nutran dichas identificaciones funcionales no han de ser sino lugares comunes. Pero resulta poco menos que evidente que pueden existir estados mentales netamente diferentes entre sí cuyas diferencias no queden recogidas en ninguna generalización que podamos justamente tildar de perogrullada; el ejemplo aducido por Block (1978: 80) –la característica astringencia desencadenada por la presencia de taninos en un vino, que la psicología popular de la percepción gustativa no parece distinguir del mero amargor– es tan bueno como cualquier otro. Ahora bien, no sólo corremos el riesgo de que los tópicos del sentido común digan bien poco sobre ciertos estados mentales –como se trasluce

en el problema de la diferenciación–, sino también de que digan demasiado sobre otros –en particular, de que lo que digan sea falso. Si los expurgamos *porque* sabemos que son falsos, estaremos adoptando en la práctica una variedad de funcionalismo empírico, pues es de suponer que sabemos tal cosa merced a alguna suerte de investigación. Queda la opción –que según Block (1978: 80) habría sugerido el propio Lewis– de intentar neutralizar esas generalizaciones falsas, pero vulgarmente aceptadas, estocásticamente, mediante una especie de cuarentena al azar en virtud de la cual fundamentaríamos nuestro análisis no en la conjunción de la totalidad de los tópicos de la psicología cotidiana, sino en una disyunción de conjunciones de la mayor parte de dichos tópicos. La táctica, innegablemente, surte efecto: es probable que para cada estado mental demos con un término de la disyunción con la que pretendemos identificarlo en el cual no aparezcan enunciados *non grati*, de modo que la verdad de ese término haría por sí sola verdadero el enunciado de identidad funcional. Basta con prolongar convenientemente la disyunción para convertir esa probabilidad en certeza –un sencillo ejercicio de matemática combinatoria. Sin embargo –como sagazmente señala Block–, el fruto apetecido se logra a costa de agravar el problema de la diferenciación, pues si ya nos es fácil hallar pares de estados mentales diferentes que son iguales con respecto a la totalidad de la psicología cotidiana, *a fortiori* lo será aun más encontrarlos si sólo han de ser iguales con respecto a la mayor parte de ella. Además –dicho sea de paso– la maniobra atribuida a Lewis acarrea una atroz merma de elegancia teórica.

Junto al problema de la diferenciación y al problema de la verdad, Block (1978: 78-79) consigna también las dificultades de la psicología cotidiana –que el funcionalismo analítico heredaría– para dar cuenta de estados mentales acontecidos en circunstancias poco frecuentes, como la parálisis, o en otras verdaderamente inusitadas –pero extrañamente habituales en el debate filosófico–, como la del cerebro mantenido con vida *in vitro* y conectado artificialmente a sus sistemas aferentes o eferentes. La ventaja crucial del funcionalismo empírico en este terreno residiría en que puede adherirse a cualquier caracterización de estímulos y respuestas que, avalada por la investigación científica, pudiera dirimir estas cuestiones –dicho de otro modo, puede trazar la linde entre el organismo y su entorno atendiendo a consideraciones teóricas. Esta posibilidad, por motivos obvios, estaría vedada para el funcionalismo analítico, que deberá plegarse a las descripciones cotidianas de estímulos y respuestas. Con todo, el problemático concepto de lugar común de la psicología cotidiana es el mejor fuste del que puede dotarse el funcionalismo analítico. Como expeditivamente señala Block (1978: 78), tratar de reemplazarlo por el de enunciado analítico no resolverá ninguno de los problemas planteados, pero suscitará tantos otros como dudas alberguemos, desde Quine (1953), respecto a la credibilidad de la distinción analítico / sintético.

La posición del funcionalismo empírico, que habitualmente se identifica con la ortodoxia cognitivista, suele modularse en términos de propiedades funcionales más que de rol causal, si bien ambos conceptos se hallan estrechamente ligados. En el marco de un severo examen de la robustez del uso que hace el cognitivismo de la

idea de representación mental, Schiffer (1986), por ejemplo, ofrece una reconstrucción sistemática de la noción de *propiedad funcional*, en la que ésta se presenta como derivada de la de rol funcional, o causal:

> A *functional role* is simply *any* second-order property of first-order state-types possession of which entails that the state-type possessing it is causally or counterfactually related in a certain way to other state-types, to outputs, to inputs, or to distal objects and their properties. […]
>
> Now, each functional role determines a unique *functional property*, viz. the property of having some property which has that functional role; since the properties which have functional roles are state-types, the functional property determined by a functional role is the property of being a token of a state-type which has that functional role. In other words, if *F* is a functional role, then the property expressed by the open sentence
>
> *x* is a token of some state-type which has *F*
>
> is a functional property. (Schiffer 1986: 128-129)

En estos términos:

> […T]he contemporary view is, then, that each belief state-type is identical to some functional property, […] whereas for each belief-property of persons there is some functional property such that the belief-property is identical to the property of having some state-token which has that functional property. (Schiffer 1986: 130)

Es decir:

> For each proposition p there is some functional role F such that being a belief that p = being a token of a state-type that has F. (Schiffer 1986: 131)

Parece que esta formulación permitiría reconstruir la idea de que las propiedades neurofisiológicas *no* son propiedades funcionales, mientras que las propiedades definidas en la tradición conductista *sí* lo son –aunque se hallen distorsionadas por una restricción metodológica artificiosa, la de atender sólo a las relaciones causales de tipos de estados internos con tipos de estímulos y de respuestas, ignorando sus relaciones con otros tipos de estados internos. En cualquier caso, el propio Schiffer señala cómo sus nociones de propiedades y roles funcionales respaldan de inmediato la tesis de realizabilidad múltiple: un tipo de estado físico de primer orden instanciaría una propiedad funcional si desempeñara el rol funcional que la determina, idénticos roles funcionales y por tanto idénticas propiedades funcionales pueden venir desempeñados por propiedades físicas heterogéneas (Schiffer 1986: 129).

A primera vista, la distinción entre propiedades y roles funcionales puede parecer superflua. No lo es –entiende Schiffer– porque la necesitamos para explicar en qué consisten propiedades tales como la que se predica de *x* en "*x* es una creencia

de que la nieve es blanca". Tales propiedades son en rigor propiedades de instancias de estados mentales (un caso particular de creencia, en el caso en cuestión). En consecuencia, no podrían identificarse con *roles* funcionales, ya que un rol funcional es una propiedad (de segundo orden) de un *tipo* de estados (Schiffer 1986: 130, *supra*). Las *propiedades* funcionales, en cambio, son propiedades de instancias: "[…] la propiedad de ser una instancia de un tipo de estados con un cierto rol funcional" (Schiffer 1986: 129)[151].

Así pues, lo que Schiffer propone es restringir el atributo "funcionalista" para la descripción de aquellas teorías de la mente, los estados mentales o las propiedades mentales según las cuales el hecho de que una propiedad mental, como la de ser una determinada creencia, se identifique con una propiedad funcional se explica a su vez de acuerdo al rol funcional que dicha creencia desempeña según una teoría psicológica. Se trata, sin embargo –Schiffer lo admite abiertamente– de una disputa puramente terminológica: hoy en día, es frecuente denominar funcionalismo a teorías que se limitan a propugnar que las propiedades de los estados mentales son propiedades funcionales, sin entrar a dirimir la cuestión de si es o no el rol funcional ocupado por cada una de ellas en la psicología del organismo lo que hace que consistan en una u otra propiedad funcional. Al fin y al cabo:

> The nebulous label "functionalism" will nowadays support nearly any antimentalist, physicalistically creditable theory of belief-properties [...] (Schiffer 1986: 130)[152]

A la disputa terminológica, con todo, subyace una cuestión de peso: la diferencia entre teorías funcionales y teorías funcionalistas. Ante un problema planteado en términos de caja negra –a saber: explicar y predecir la conducta de un sistema conociendo la estimulación que recibe pero no sus estados internos–, una *teoría funcional*, a juicio de Schiffer (1986: 131-132), no es más que una solución tentativa que introduce como entidades teóricas determinados estados internos del sistema definidos según sus relaciones entre sí, con los estímulos y con las respuestas; se

[151] Aclarar la naturaleza de las *propiedades* que se predican de los estados mentales –ya entendidos como tipos, ya como instancias concretas– es, según el planteamiento de Schiffer, una parte crucial del problema mente-cuerpo. Las otras dos son: aclarar la naturaleza de los *estados* mentales concretos, y en particular su relación con los estados físicos del organismo, y aclarar la naturaleza de las mentes, entendidas como aquellas *entidades* que satisfacen oraciones abiertas tales como "*x* siente dolor". Aunque su argumento queda ileso, despunta en Schiffer el error que tanto Aristóteles como Tomás, y luego Wittgenstein, se afanaron en denunciar: no es mi mente la que sufre o piensa, sino yo –*anima mea non est ego*.

[152] Como es obvio, Schiffer está empleando el término "antimentalista" en un sentido muy diferente del instaurado por Fodor (1968b), según el cual el funcionalismo es rotundamente una posición mentalista. Por lo demás, su exigencia de una determinación funcional de la identificación de propiedades de estados mentales con propiedades funcionales parece equivaler precisamente a la que abandonaría Fodor al renunciar al internismo y asumir, casi como un mal irremediable, la necesidad de una semántica externalista, inspirada en la noción de información.

trata, pues, de una teoría empírica. Una teoría funcionalista, desde luego, es desde este punto de vista algo muy diferente:

> Functionalism is a philosophical theory about the nature of propositional attitudes […]. What makes a functionalist a functionalist is the way he explicates propositional attitudes in terms of functional theories.
>
> The functionalist […] holds that
>
> Some psychological theory determines a correlation of each proposition *p* with a functional role indexed by *p* in that theory in such a way that being a belief that *p* = the functional property of being a token of some (first-order, physical) state-type that has that functional role. (Schiffer 1986: 134)

La cuestión inmediata, desde luego, sería qué teoría psicológica determina la correlación de cada proposición *p* con un rol funcional adecuado; la respuesta a esa pregunta permite perfilar de nuevo las dos variedades cardinales del funcionalismo, aplicadas esta vez no ya a la tarea de determinar, digamos, qué hace de *x* una creencia sino que hace de *x* una creencia de que *p*. Nos encontraríamos así, una vez más, bien ante un funcionalismo analítico o de sentido común, según el cual la teoría que establece esa correlación es la psicología popular o *folk*, bien ante un funcionalismo empírico o psicofuncionalismo, según el cual el mayorazgo de los lazos entre el contenido de los estados mentales y la función de los estados fisiológicos de los organismos pertenece a la psicología cognitiva –o, en cualquier caso, a una psicología científica madura.

La proposición, entonces, sirve de índice en la teoría psicológica para el rol funcional de un tipo de estados físicos cuyas instanciaciones poseen la propiedad funcional en la que consiste el hecho de ser una determinada actitud proposicional. La idea es que, en el funcionalismo, las proposiciones se adhieren a una teoría funcional como claves o índices externos cuyas relaciones reflejan las que se dan entre los estados internos del sistema:

> Then we might hope to ascribe functional roles to internal state-types via a quantification over functions that map propositions onto state-types whose causal relations to one another, to inputs, and to outputs mirror the relevant logical (or other) relations in which their correlated propositions stand to other propositions. […]
>
> […] So this […] is how propositions might enter into a functional theory: as objects wholly external to the system and its workings to which we refer in order to enable us to ascribe functional roles to unknown physical state-types of the system –i.e., unavailable state-types that enter into unavailable causal laws that are explanatory of the system's behavior at a deeper level than that to which the functional theory aspires. (Schiffer 1986: 132-133)

Como habría señalado Stalnaker (1984), el papel de las proposiciones en la teoría psicológica sería así –después de todo– análogo al de los números en la teoría física. Poco después consignaría Cummins (1989), en su notable estudio del papel del

concepto de representación en la explicación psicológica, que la analogía puede dotarse de un alcance mucho más amplio[153]:

> The concept of representation invoked by the C[omputational] T[heory of] C[ognition] is the same concept that is implicit in the sort of mathematical science that Galileo invented. It is the sense in which a graph or equation represents a set of data, […], a parabola represents the trajectory of a projectile, intelligence cannot be represented on a ratio scale […], and we ask whether social or economic dynamics can be adequately represented by a set of linear equations […]. (Cummins 1989: 96-97)

A un sentido si se quiere menos exigente de la idea de funcionalismo que el bosquejado por Schiffer (1986) parece apelar Block (1980a: 27) cuando apunta que bien puede entenderse por funcionalismo sencillamente una estrategia de investigación orientada a la construcción de *explicaciones funcionales*: explicaciones que den cuenta del comportamiento de un sistema apelando a las características de las partes que lo integran y el modo en que están organizadas. Éste sería el esquema de la explicación por "análisis funcional" que con todo detalle describiera Cummins (1975). Ahora bien, no hay menoscabo alguno de su legitimidad o su valor en afirmar que, por razones que venimos examinando con cierto detenimiento, la idea de análisis funcional difícilmente habría llegado a alterar los derroteros históricos de la psicología científica de no haberse visto ligada a la de computación, con ello a la de representación, y con ello al esbozo de una fundamentación sólida del empleo de conceptos intencionales en la teorización psicológica. En todo caso, como el propio Block (1978: 81, *infra*) había dejado ya apuntado, sería infructuoso erguir una defensa *a priori* del funcionalismo mediante el expediente de recortar sus compromisos a ese *minimo minimorum*, y pretender entonces que, así rebajados, son comunes a toda ciencia.

Cabría, no obstante, una idea de funcionalismo con implicaciones más firmes que la de una mera estrategia orientada a la construcción de explicaciones funcionales en el sentido desglosado por Block, o de teorías funcionales en el sentido de Schiffer, pero desprovista de compromisos de orden ontológico. Lo que Block denomina "funcionalismo computacional-representacional" podría entenderse como una tesis acerca de la naturaleza de la explicación científica en psicología, pero en la medida en que sea parte de esa tesis la necesidad de incorporar al vocabulario psicológico los conceptos de computación y representación (es decir, de computación sobre representaciones), sería ya una tesis que desbordaría la presunta relación de identidad entre explicación científica y explicación funcional: desde luego, no de toda explicación que apele a la organización de las partes de un sistema para rendir cuentas de su comportamiento puede decirse que apele a computaciones sobre

[153] Una interesantísima reflexión sobre el concepto de representación desde una perspectiva cercana a ésta puede encontrarse en Ibarra (2000), donde se intenta construir una visión pluralista y pragmática, basada en el mínimo común de su naturaleza vicaria y en la noción de razonamiento subrogatorio, que permita abarcar los "muchos rostros" (Ibarra 2000: 38) que a lo largo de la historia del pensamiento nos muestra la idea de representación.

representaciones. Aún así, la transición desde ese "funcionalismo computacional-representacional" hacia lo que el mismo Block denomina "funcionalismo metafísico" era poco menos que inevitable: si articulamos un proyecto de explicación de un determinado conjunto de fenómenos sobre la estrategia de tratarlos como si se ajustaran a determinadas especificaciones, y encontramos que el proyecto tiene perspectivas de prosperar, es difícil resistirse a la tentación de considerar que la naturaleza de los fenómenos en cuestión se ajusta de hecho a las especificaciones desplegadas. Si considerar los fenómenos psicológicos como computaciones sobre representaciones nos permite explicarlos convincentemente, sólo un ánimo de disciplinadísima frugalidad filosófica podría renunciar a plantearse que los fenómenos psicológicos sean, en efecto, precisamente eso: computaciones sobre representaciones –aunque eso no excluya que, desde otras perspectivas, puedan acaso ser también otras cosas. Una renuncia así no es, desde luego, indefendible –ya se ha dicho: el paso hacia una tesis de orden metafísico es *casi* inevitable, no inevitable–, pero como anota Fodor (1968: 186-187) al hilo de una controversia estrechamente ligada a ésta, "[…] éste es el punto en el que las cuestiones filosóficas acerca de la psicología quedan absorbidas por cuestiones filosóficas más generales acerca de las teorías científicas" –en particular, por el debate en torno al realismo científico. Es difícil, ciertamente, no convenir con tal conclusión:

> "Supongamos que tenemos una teoría psicológica que postula P, y que tenemos un buen fundamento para considerarla como verdadera. ¿Estará, por ello, justificado que atribuyamos P al organismo cuya conducta pretende explicar esta teoría?"
>
> […] A mi entender, todas las razones que se pudieran dar para negarse a atribuir P a un organismo en las condiciones anteriores, serían las mismas que llevaran a pensar que *todas* las entidades teoréticas son ficciones o meros constructos lógicos. (Fodor 1968: 186-187)

Dicho de otro modo, si la hipótesis de que los procesos psicológicos son computaciones sobre representaciones nos permite construir explicaciones convincentes de la conducta de los organismos –y, claro, de ciertas máquinas–, entonces los motivos para rechazar la existencia de procesos psicológicos se reducen a los que pudiéramos tener para rechazar la existencia de los referentes de los constructos teóricos de cualquier teoría científica, por acreditada que estuviera su verdad.

También Bechtel (1988: 152-153) reconoce en Lewis (1972, 1980) una versión del proyecto funcionalista que toma por fundamento el conocimiento psicológico de sentido común. En cambio, erigir una concepción funcionalista de la mente sobre los planos proporcionados por la labor científica habría sido –a su entender– el propósito compartido en los influyentes trabajos de Putnam (1967a, 1967b), Fodor (1968b, 1975, 1980a, 1983) y Dennett (1981, 1987, 1991), a cuyas formas de plantear el funcionalismo denomina Bechtel respectivamente "funcionalismo de tabla de máquina", "funcionalismo computacional", y "funcionalismo homuncular".

Las diferencias entre la concepción de lo mental que Putnam propugnaba a finales de los años sesenta –en uno de cuyos más feroces críticos habría de convertirse– y la que Fodor –pese a un buen número de rectificaciones de mayor o menor calado– ha venido defendiendo desde entonces pueden considerarse menores. Pero entre Fodor y Dennett parecen mediar discordancias tan estrepitosas que la mera adscripción de ambos a idéntico corrillo teórico puede resultar insólita. Es oportuno recordar, pues, cuán exigua es la coincidencia que se les atribuye: se trata tan sólo de la idea de que es el desarrollo científico de la psicología quien debe dictar los pormenores de nuestras nociones filosóficas acerca de lo mental.

El intento de convertir esta concepción mínima del funcionalismo empírico en un argumento *a priori* que lo respalde ha sido identificado por Block (1978), que lo considera falaz. La estrategia pasaría por postular como principio general de nuestra teoría de la ciencia que responder acerca de la naturaleza de los fenómenos que estudia es competencia de cada rama de la ciencia; la respuesta de la psicología respecto a la naturaleza de los estados mentales sería que son estados funcionales, y el funcionalismo empírico se limitaría a dar por buena esa respuesta en virtud del principio general mencionado. Visto desde esa perspectiva, en efecto,

> […p]sychofunctionalism is just the doctrine that mental states are the "psychological states" it is the business of psychology to characterize. (Block 1978: 81)

El argumento descarrila, a juicio de Block, tan pronto como advertimos que ni siquiera es viable, de hecho, en el caso de la física. Dado que las denominadas entidades "duales" –como protones y antiprotones, electrones y positrones, etc.– son, hasta donde sabemos, funcionalmente idénticas excepto en sus relaciones causales recíprocas, una teoría que las identificara con propiedades funcionales expresadas en el formalismo de Ramsey estaría condenada a confundirlas. Ahora bien, lo que esto muestra es, como mucho, que la física no respondería en términos funcionales a la pregunta por la naturaleza de las entidades duales, y eso no contradice ninguna de las premisas del argumento esbozado por Block –en particular, no contradice la premisa de que la psicología sí lo hace. Completar esta línea de razonamiento exigiría a Block encontrar un análogo mental a las entidades duales de la física subatómica, tarea que ni siquiera acomete. Del mismo modo, lo hace deudor de la pervivencia teórica de la estricta simetría entre la materia y la antimateria, que bien puede quedar en entredicho en cualquier momento –de hecho, la despareja presencia de ambas en el universo, una de las incógnitas más longevas de la física cuántica, acaso apunte precisamente en esa dirección. Además, incluso si lograra salvar estas dificultades, Block debería vérselas con la objeción de que la incapacidad de rendir cuentas de (tipos de) entidades (propiedades, estados) que se diferencien funcionalmente entre sí sólo en sus relaciones recíprocas no es más que una deficiencia del formalismo de Ramsey, que como tal habrá de ser subsanada mediante refinamientos formales. Entendido, pues, como la conjunción de la tesis según la cual compete a la psicología científica determinar la naturaleza de lo mental –en general: a cada rama de la

ciencia, de su objeto de estudio– y la apuesta por una psicología de corte funcional, el funcionalismo empírico es más robusto de lo que Block parece dispuesto a conceder. Desde luego, su pujanza depende de los logros explicativos del programa de investigación funcionalista en psicología, pero no de los que pudieran dejar de cosechar programas análogos en otras áreas de la ciencia[154].

A renglón seguido ensaya y refuta Block otra falacia esgrimida en favor del funcionalismo empírico: que los estados mentales son estados psicofuncionales sencillamente porque el psicofuncionalismo proporciona la mejor explicación disponible de la naturaleza de lo mental. Es el mismo patrón de inferencia hipotética, de raíz abductiva, que Dennett (1991a: 329, *infra*) ha impugnado en relación con la hipótesis del lenguaje del pensamiento: "¿qué otra explicación podría haber?". Como señala Block (1978: 82), inferencias de esta ralea no son admisibles cuando la presunta explicación óptima es claramente deficiente, y menos aun cuando ni siquiera tenemos la seguridad de que *haya* una explicación que satisfaga nuestra pregunta, dado que cabe albergar sospechas de que ésta no haya sido correctamente formulada. Además, Dennett sin duda se encargaría de recalcar –fiel a su costumbre– que nuestra incapacidad de concebir explicaciones distintas de tal o cual fenómeno seguramente no revele más que la pobreza de nuestra imaginación.

La conclusión de estas escaramuzas es, en fin, que el psicofuncionalismo no puede sino plegarse a su condición al menos parcialmente empírica, o, dicho de otro modo, a su compromiso con la vigencia de alguna taxonomía funcionalista de los estados internos en la explicación científica de la conducta humana. Uno tras otro, los diversos intentos de revestir al funcionalismo empírico de un carácter apodíctico fracasan irremediablemente, como ya sucediera a Hempel (1935) en su afán de hacer lo propio con el conductismo.

Que el funcionalismo debe entenderse como una tesis de carácter netamente empírico era una idea central –aunque sujeta a los muchos matices cuyo desentrañamiento hemos acometido *supra*– en los trabajos seminales de Putnam (1960, 1963a, 1967a, 1967b). No obstante, Fodor pronto comenzaría, de la mano de Block, a distanciarse de algunos planteamientos de Putnam. En el mismo trabajo en el que descartaran la viabilidad de una reconstrucción disyuntiva del conductismo lógico o el fisicalismo de tipos, Block y Fodor (1972a) exponían también sus reservas acerca del "funcionalismo de tabla de máquina", que denominaban entonces "teoría de identidad de estados funcionales", o *FSIT:* "Functional State Identity Theory". La querella giraba, de nuevo, en torno a la naturaleza de las propiedades cuya articulación en tipos de estados que las poseen hubiera de permitirnos reflejar una taxonomía verosímil y explicativamente fructífera de los estados psicológicos; la

[154] Distinto sería –como hace Dennett (2001)– plantear que la caracterización funcionalista de los fenómenos es un rasgo general del saber científico, y reemplazar con esa tesis la tímida apuesta respecto a que la caracterización de los estados mentales en que finalmente se asiente, en su madurez, la psicología científica será funcionalista. Para ese argumento sí sería nociva la objeción de Block, si bien quedaría al albur de que pequeñas correcciones del canon formal adoptado por el funcionalismo la tornaran del todo inofensiva.

solidez de la noción de máquina de Turing como formalismo para la descripción global del sistema cognitivo quedaba al margen de una disputa, por lo demás, muy mitigada –dado que el propio Putnam había revisado el artículo. En efecto, la doble conclusión de Block y Fodor era que:

> It may be both true and important that organisms are probabilistic automata. But even if it is true and important, the fact that organisms are probabilistic automata seems to have little or nothing to do with the conditions on type identity of their psychological states. (Block y Fodor 1972a: 60)

Es decir: donde el funcionalismo de tabla de máquina se torcía era en su pretensión de asignar una única descripción como autómata probabilista a cada sistema capaz de albergar estados mentales de tal manera que cada uno de sus estados mentales quedara identificado con uno de los estados de tabla de máquina especificados en la descripción. Salvando esa ambición –que, al margen del desengaño de uno tras otro intento de colmarla, Block (2007b) habría de escudar como la legítima e irrenunciable vocación metafísica que alienta nuestra concepción de la mente–, el funcionalismo de tabla de máquina aventajaba holgadamente tanto al conductismo como al fisicalismo. En particular, nos permitía incluir sus vínculos recíprocos en la caracterización de los estados psicológicos, arrojando luz sobre lo que en el conductismo eran tinieblas: "el inconfundible carácter indirecto de la relación entre estados psicológicos y conductas", o el hecho de que "[…] estados psicológicos que carecen de *toda* expresión conductual puedan pese a ello ser distintos" (Block y Fodor 1972a: 50). Además, merced a su compatibilidad con el fisicalismo de casos y, con ello, con la postulación de relaciones causales entre (instancias de) estados psicológicos, el funcionalismo de tabla de máquina acomodaba sin dificultad cualquier relación causal entre estados del organismo que una teoría de corte fisicalista hubiera podido establecer (sencillamente, incorporándola a la tabla de máquina) –al igual, por supuesto, que cualquier relación causal entre estímulos y respuestas establecida por la investigación conductual. Así que, a fin de cuentas, "[…] any advantages which accrue to causal analyses of the psychological states, or of the relations between psychological states and behavior, equally accrue to *FSIT*" (Block y Fodor 1972a: 50).

Éste es –claro está– el trasfondo de la ecuanimidad con la que Fodor (1981a: 9, *supra*) tasaba la valía del conductismo y el fisicalismo como concepciones de lo mental en, respectivamente, su apreciación del carácter relacional y del carácter autónomo –*ergo* causalmente eficiente– de los estados psicológicos, virtudes todas que el funcionalismo habría heredado de sus ancestros filosóficos.

Recortado, pues, el alcance del ataque, Block y Fodor (1972a: 50-58) revisan hasta seis peculiaridades de nuestros estados mentales que a su juicio inhabilitan al funcionalismo de tabla de máquina como proveedor de una tipología naturalizada y verídica de dichos estados: (*i*) la distinción –ryleana, después de todo– entre *estados psicológicos disposicionales* y *actuales*, (*ii*) la *intervención simultánea* de varios estados psicológicos en la determinación de la conducta, (*iii*) el *carácter cualitativo* de ciertos

estados mentales, (*iv*) el modo en que la clasificación cotidiana de estados mentales en tipos ignora *diferencias irrelevantes*, (*v*) la *productividad* inherente a muchos estados psicológicos en virtud de su carácter proposicional, y (*vi*) la *sistematicidad* que, por idéntica razón, también parece serles consustancial. Puesto que se trata de los hitos principales del sendero que condujo al temprano abandono del funcionalismo de tabla de máquina en favor de lo que se dio en llamar "funcionalismo computacional", conviene estudiarlos en detalle.

Digamos –por dar al litigio conceptual la concreción de un ejemplo cotidiano– que el sujeto cuya psicología tratamos de diseccionar habla nuestro idioma, cree que va a seguir lloviznando toda la tarde y se halla envuelto en una vaga pesadumbre; también cree que la lluvia, en su debido tiempo y cuantía, es buena para las cosechas. La primera tara que Block y Fodor (1972a: 51) encuentran en el funcionalismo de tabla de máquina sería (*i*) su incapacidad de dar cuenta de la diferencia entre fenómenos psicológicos consistentes en una disposición, como hablar un idioma, y fenómenos psicológicos consistentes en un estado actual, como hallarse envuelto en una vaga pesadumbre –o, según su ejemplo, estar hablando un idioma. Por supuesto, el funcionalismo de tabla de máquina podría recurrir a describir las disposiciones como estados de tabla de máquina de un autómata probabilista, y los estados actuales como el estado de tabla de máquina en el cual se encuentra el autómata. Pero eso constituiría –argumentan Block y Fodor– un esfuerzo inútil: mientras que un autómata probabilista puede en principio llegar a estar en cualquiera de los estados definidos en su tabla de máquina, no es cierto que nosotros podamos atravesar actualmente cualquiera de los estados mentales que albergamos como disposiciones.

Sin embargo, la conclusión de Block y Fodor se basa en un análisis tan precipitado como cándido de las relaciones entre una potencia y su acto. A su entender, a los predicados disposicionales que describen una capacidad, como "habla nuestro idioma" corresponden los predicados actuales que describen su ejercicio, como "está hablando nuestro idioma", y a los predicados disposicionales que describen un rasgo, como "es avaricioso" les corresponden los que describen algo así como su manifestación, como "está siendo avaricioso"; pero los predicados disposicionales que describen una actitud proposicional, como "cree que *P*" o "desea que *Q*", carecen de vertiente actual: sería absurdo –según eso– decir de alguien que "está creyendo que *P*" o que "está deseando que *Q*". Ahora bien, todo esto se halla muy lejos de conformar la obviedad que Block y Fodor parecen creer. Ante todo, es obvio que los actos reseñados –estar hablando un idioma, mostrarse avaricioso– no son, bajo esa descripción al menos, estados mentales actuales, sino más bien conductas o conjuntos de conductas. De hecho, es más sencillo dar con pares de potencia y acto de estados mentales precisamente en el terreno de las actitudes proposicionales, del que Block y Fodor pretenden expulsarlos: en el caso de nuestro melancólico sujeto y de su vida mental reciente, por ejemplo, parece sensato interpretar –es, al menos, una interpretación posible– que su creencia de que va a

seguir lloviznando toda la tarde habrá sido una creencia actual, algo que efectivamente ha creído en un momento dado –"ha pensado", diríamos–, mientras que su creencia de que la lluvia es, en su debido tiempo y cuantía, buena para las cosechas resulta ser, en el contexto de su vida mental reciente, una creencia meramente disposicional, que sólo podemos fijar, a la manera ryleana, mediante enunciados hipotéticos acerca de cómo respondería el sujeto a determinados estímulos –fundamentalmente, a la pregunta de si la lluvia es buena para las cosechas. Por supuesto, otras creencias de nuestro sujeto acaso sean meramente disposicionales en el contexto de la *totalidad* de su vida mental pasada –tal vez la creencia de que bajo la misma lluvia coexisten ráfagas que caen en diferentes ángulos respecto a la superficie terrestre, o a la perpendicular aproximada que su cuerpo traza con ésta cuando está de pie…–, y algunas, sencillamente, no serán atribuibles a nuestro sujeto, ni siquiera disposicionalmente –la creencia de que la lluvia es un don de los dioses, o la de que cada una de sus gotas está poblada por una multitud de criaturas prodigiosas. De acuerdo, pero nada de eso entraña ni remotamente que haya alguna creencia que podamos albergar como disposición pero no atravesar como estado, alguna creencia cuya potencia adolezca de acto –*idem*, *mutatis mutandis*, respecto de los deseos, o de cualquier otra actitud proposicional. La mera idea de que podamos cobrar consciencia de una actitud proposicional parece de hecho entrañar la de que ésta se nos haga *presente* como un estado actual[155]. Así pues, la convicción que muestran Block y Fodor de que existen disposiciones de esa índole –a saber, disposiciones que no pueden tornarse en acto– asoma como un inesperado residuo de conductismo lógico en su concepción de la realidad mental. Dado que descansa sobre ese convencimiento, la conclusión de que el funcionalismo de tabla de máquina no puede dar cabida a la diferencia entre estados mentales disposicionales y actuales *porque, a diferencia de* un sistema cognitivo humano, un autómata probabilista puede en principio atravesar todos los estados de su tabla de máquina, queda por lo tanto

[155] Es más bien la intuición contraria –que todo estado que intuitivamente pudiéramos considerar intencional ha de poder en principio tornarse actual aflorando a la consciencia– la que en ocasiones se ha apuntado para señalar la divergencia entre, por un lado, las representaciones de la ciencia cognitiva –al menos en tanto se entiendan éstas, en la estela de Chomsky, como el ámbito de aplicación de las reglas que han de servirnos de modelo de alguna competencia cognitiva– y, por otro lado, la noción de representación que se hallaría incorporada al discurso ordinario sobre las actitudes proposicionales; es el caso, por ejemplo, de Searle (1989, 1990a, 1992) y la tesis que él denomina "Principio de Conexión".

Tan es así, que un estudioso del funcionalismo de la talla de Bechtel, cuando resume las críticas que Block y Fodor dirigieron contra el funcionalismo de tabla de máquina, presenta esta objeción en particular mediante el ejemplo, precisamente, de una creencia: el funcionalismo de tabla de máquina, según la lectura que de Block y Fodor plantea Bechtel, "[…] no puede captar la importante distinción entre estados mentales que ocurren efectivamente (contemplar efectivamente la proposición de que, si hay nubes que amenazan lluvia y truenos, entonces la lluvia vendrá a continuación) y estados disposicionales (creer pero no contemplar activamente la proposición de que, si hay nubes que amenazan lluvia y truenos, entonces la lluvia vendrá a continuación)" (Bechtel 1988: 156). La posición de Block y Fodor es en este punto insólita en grado tal que Bechtel parece sencillamente haberla *corregido* sin darse cuenta.

gravemente debilitada. Como mucho, la dificultad para dar cuenta de la distinción entre estados mentales disposicionales y actuales que aplicamos en nuestro propio caso se cifrará en la dificultad para dar cuenta del papel que desempeñe la consciencia en estos últimos: no es poco, pero tampoco es, por supuesto, una particularidad del funcionalismo de tabla de máquina[156].

Más alarmante resulta quizá la segunda vía de agua que Block y Fodor advierten en el casco del funcionalismo de tabla de máquina. Se trata de la incapacidad del modelo del autómata probabilista para hacerse, por ejemplo, con el hecho de que la creencia de que seguirá lloviznando toda la tarde *y* la vaga pesadumbre que envuelve a nuestro sujeto puedan conjuntamente motivar una conducta –tal vez tararear cierta canción– que ni la creencia ni el sentimiento engendrarían por sí solos[157]. Ciertamente, (*ii*) la ocurrencia simultánea de más de un estado psicológico resulta difícilmente trasladable a la descripción de un autómata en términos de su tabla de máquina:

> […] *FSIT* can provide for the representation of sequential interactions between psychological states, but not for simultaneous interactions. Indeed *FSIT* even fails to account for the fact that an organism can be in more than one ocurrent psychological state at a time, since a probabilistic automaton can be in only one machine table state at a time. (Block y Fodor 1972a: 53)

Es crucial apreciar que tanto la simultaneidad de estados psicológicos actuales como la concurrencia de causas mentales sobre un mismo efecto se nos antojan intratables. La tabla de máquina bien puede especificar, por ejemplo, que la conducta O_3 se emitirá si se da el estado S_3, el cual se produce si coinciden el estímulo I_1 y el estado S_1: un estímulo y un estado interno concurren así como causas de una conducta. Pero no cabe consignar en una tabla de máquina que S_3 se origine al confluir S_1 y S_2, porque una máquina de Turing es un dispositivo serial, una arquitectura

---

156 No deja de ser curioso, por otra parte, que Block y Fodor (1972: 52) den por sentado, sin argumentarlo, que el acto de creer que *P* tendría que ser, al igual que su potencia, de naturaleza inconsciente, acaso porque asuman que la explicación de los estados mentales conscientes queda demasiado alejado de las facultades del funcionalismo –en lo cual, por cierto, coincidirían con Searle (1980). Pero incluso la distinción entre consciencia fenoménica y consciencia de acceso, sobre la que Block (1995) edifica su propio concepto de la función de la consciencia, resultaría difícilmente sostenible si no se respaldara la noción de consciencia de acceso mediante alguna suerte de distinción entre actitudes proposicionales disposicionales y actuales.

157 Por cierto, resulta sorprendente que el ejemplo de ocurrencia simultánea de dos estados mentales al que recurren Block y Fodor (1972: 53) sea precisamente el de la coincidencia de una emoción y un pensamiento. Es difícil admitir que "cree que *P*" sea un estado disposicional que carece de faceta actual, que "está pensando que *P*" sea un estado mental actual, que puede concurrir en el tiempo con otros, y que la relación entre ambos no sea tal que "está pensando que *P*" constituye la faceta actual de "cree que *P*". Más sensato parece concluir que usamos distintos verbos para describir el mismo estado mental sólo en forma de disposición –"cree que *P*", pero no *"está creyendo que *P*"– o tanto en forma de disposición como de acto –"piensa que *P*", "está pensando que *P*".

computacional tipo von Neumann que recorre sus estados posibles de uno en uno[158]. Eso no deja prácticamente otra opción que reconocer –como hacen Block y Fodor– que:

> […] if probabilistic automata are to be used as models of an organism, the appropiate model will be a set of intercommunicating automata operating in parallel. (Block y Fodor 1972a: 53)

La cuestión de la simultaneidad de estados psicológicos actuales queda, a simple vista, razonablemente resuelta con la postulación de autómatas gregarios: el organismo atraviesa el estado S_1 y el estado S_2 a la vez cuando uno de los autómatas –digamos T_1– se encuentra en S_1 y otro, T_2, en S_2. Pero la concurrencia de estados mentales como causas de una conducta, o de otro estado mental, se muestra algo más refractaria a esta táctica. Conviene advertir que Block y Fodor definen los miembros del conjunto que sirve de modelo del organismo como autómatas *intercomunicantes*. Es obvio que sin esa capacidad de intercomunicación no habrá posibilidad de que los estados internos actuales de más de un autómata confluyan como causa del estado interno de uno de ellos, o de otro, o de una respuesta de uno de ellos, o de otro. Pero conferir a los autómatas dicha capacidad –evitar que funcionen en completa ignorancia unos de otros– requiere que las tablas de máquina de cada uno de ellos puedan incluir referencias a estados internos de otros, las cuales incorporen además alguna clave que identifique al autómata a cuyos estados internos apela dicha referencia. De lo contrario, tendríamos que dotar al sistema de algún mecanismo externo al conjunto de autómatas, encargado de regir las interacciones entre ellos. Sea cual sea el camino que elijamos –acoplar a las propias tablas de máquina de nuestros autómatas gregarios las especificaciones relativas a sus relaciones recíprocas, o externalizarlas guarneciendo al sistema de una suerte de pastor–, nos habremos apartado significativamente del punto de partida. La tesis de que un (tipo de) estado psicológico es (idéntica a) un (tipo de) estado de tabla de máquina de un autómata probabilista emerge ahora –en un juicio indulgente– como una verdad a medias, ya sea porque omite que el estado de tabla de máquina en cuestión puede incluir referencias a estados de tabla de máquina de otros autómatas, ya porque omite que el autómata forma parte de un sistema gobernado por otra tabla de máquina, la que describe el funcionamiento del pastor, de la que no hay ni rastro en la tabla de máquina del autómata con uno de cuyos (tipos de) estados estamos identificando el (tipo de) estado psicológico que pretendemos explicar. Si, por último, descartamos

[158] Cabe pensar, es cierto, en una máquina de Turing no determinista como un sistema que se ramifica en un árbol computacional, examinando las consecuencias de todas sus conductas posibles (pares <eferencia – nuevo estado interno>) para cada par <aferencia – estado interno> que admita más de una. Es obvio que así entendida una máquina de Turing no es una arquitectura serial, si bien puede simularse mediante una por la misma razón que cualquier máquina de Turing no determinista puede simularse, con incremento del tiempo de computación, por medio de una determinista, y su funcionamiento pueda ser remedado por varias arquitecturas seriales –autómatas gregarios, como se apuntará a renglón seguido.

ambas opciones y nos inclinamos por redescribir el conjunto de autómatas como un autómata único –un ejercicio cuya viabilidad queda garantizada por la propia noción de máquina de Turing–, acaso recuperemos la idoneidad de la identificación de los estados psicológicos globales del organismo con estados de tabla de máquina de *ese* autómata, pero lo haremos dilapidando la herramienta de descripción de estados psicológicos simultáneos que habíamos ingeniado. La estrategia más atractiva desde el punto de vista de la teorización psicológica sería –por supuesto– concedernos la libertad de alternar esos niveles de descripción –el autómata global, los autómatas gregarios– en función de nuestros objetivos explicativos puntuales. Pero eso, una vez más, obligaría a una reformulación sustancial de la tesis de identidad de (tipos de) estados psicológicos y (tipos de) estados de tabla de máquina, so pena de convertirla en una verdad parcial o un mero subterfugio. El veredicto de Block y Fodor (1972a: 53) es que la postulación de autómatas gregarios resuelve el problema de la concurrencia causal de estados internos simultáneos sin dañar la "tesis principal" del funcionalismo de tabla de máquina: la identidad entre los estados psicológicos de un organismo y los estados de tabla de máquina de su descripción óptima como autómata probabilista. Pero, por las razones que acabamos de ver, ese veredicto resulta a todas luces demasiado benévolo.

También se halla dañado el armazón del funcionalismo de tabla de máquina, según Block y Fodor, en lo concerniente (*iii*) al carácter cualitativo que tiñe ciertos estados mentales. Resulta obvio que si un autómata se encontrara en un estado de tabla de máquina que –suponiendo sorteados los últimos escollos– mostrara exactamente las mismas relaciones con estímulos, respuestas y otros estados internos que caracterizan a la vaga pesadumbre de sujeto de nuestra imaginaria investigación, y a pesar de ello no *sintiera* una vaga pesadumbre, un aspecto cardinal de la naturaleza de la pesadumbre habría quedado excluido de la concepción funcionalista de los estados mentales; lo mismo podría decirse respecto del dolor, la alegría o los celos, del olor a hierba recién cortada o el sabor del vinagre. Pensar en un autómata nos instiga a interpretar la afirmación de que éste no sintiera dolor como un mero circunloquio para atestiguar que no sentiría *nada*. Pero si el ejemplo se plantea en términos de criaturas cuya organización biológica resultara muy diferente de la nuestra –acaso cuyo ecosistema lo fuese también–, entonces se vuelve más asequible la idea de que –pese a que su estado mental de dolor, alegría, o sabor a vinagre quedara óptimamente descrito por relaciones funcionales idénticas a las que caracterizan los nuestros, es decir, pese a que viniera descrito como *el mismo estado de tabla de máquina*–, tal criatura pudiera no sentir dolor *sino* otra cosa, alegría sino quién sabe qué, o un sabor igual al que en nosotros evoca el vinagre, sino acaso las frambuesas[159]. Ahora bien: resulta obvio que si alguna de estas cosas fuera cierta,

[159] Fue de hecho en este trabajo –como recuerda Block (2007b: 11)– donde la posibilidad de concebir organismos cuyos estados mentales carecieran de carácter cualitativo pese a ser funcionalmente indistinguibles de los nuestros quedó bautizada como el problema de los *qualia ausentes* –con el tiempo, sobre todo a partir de Dennett (1991) y Chalmers (1996), fue extendiéndose la broma de

daríamos por incompleta la concepción funcionalista de lo mental; quizá sea obvio incluso que bastaría con que pudieran ser cierta para que debiéramos dictaminar la insuficiencia de la explicación funcionalista, pero no parece ser una obviedad ni que tales circunstancias sean posibles, ni que sean siquiera concebibles, ni –desde luego– que del hecho de que fueran concebibles pudiéramos deducir que fueran posibles. Eso sí: también parece obvio que sea cual fuere la resolución de una controversia tan espinosa afectaría por igual al funcionalismo de tabla de máquina que a cualquier otra variedad de funcionalismo –si bajo ese epígrafe abarcamos a todas aquellas teorías que identifiquen los tipos de estados mentales con tipos de estados definidos por sus relaciones con estímulos, respuestas y otros estados delimitados del mismo modo. Como anota Levin:

> Functionalist theories of all varieties –whether analytic or empirical, FSIT or functional specification– attempt to characterize mental states exclusively in relational, specifically causal, terms. A common and persistent objection, however, is that no such characterizations can capture the qualitative character, or "qualia", of experiential states such as perceptions, emotions, and bodily sensations, since they would leave out certain of their essential properties, namely, "what it's like" (Nagel, 1975) [Nagel 1974] to have them. (Levin 2004: §5.1)

Si el carácter cualitativo se muestra refractario al análisis funcionalista, *a fortiori* habrá de hacerlo ante el análisis conductual: nada hace pensar que una merma de las claves que usemos para capturar cierto tipo de estado mental pueda servirnos para taxonomizarlo con mayor delicadeza. Aunque –como se ha visto– la teoría de la identidad psicofísica germinó en el pensamiento de Place (1956), en parte, como un intento de reparar precisamente esa tara del conductismo, entregando el análisis en términos de procesos cerebrales de unas sensaciones crudas ("*raw feels*") que, como el dolor, no se plegaban mansamente al caudillaje de Ryle a modo de posdata que culminaría su programa reduccionista, lo cierto es que tan concebible –o inconcebible– resulta a primera vista que los estados mentales de dos organismos funcionalmente indistinguibles difieran en carácter cualitativo como que tal cosa suceda entre dos organismos fisiólogicamente indistinguibles. Si hay un enigma que gire en torno a cómo es sentir algo, se trata de un enigma que empaña cualquier concepción de lo mental que evite postular el carácter cualitativo como una de sus propiedades básicas, irreducibles.

Aun así, la idea de que el fisicalismo –en particular, la teoría de identidad psicofísica– esté más preparado que el funcionalismo para ofrecer una explicación de las sensaciones crudas en lo que atañe a su aspecto fenoménico, o, más en general, de la consciencia o la subjetividad, se halla notablemente extendida en la discusión

---

denominar *zombies* a tales organismos hipotéticos–, y también donde la posibilidad de concebir organismos cuyos estados mentales vinieran revestidos de carácteres cualitativos sistemáticamente diferentes de los nuestros empezó a conocerse como el problema de los *qualia invertidos* –aunque para el caso más debatido, el de la experiencia de color, sigue siendo habitual hablar del problema del *espectro invertido*.

contemporánea, si bien rara vez es objeto de un mínimo escrutinio. Tal vez una de las fuentes de tal convicción se encuentre en Feigl (1958: 451), quien descartaba la adscripción de sensaciones conscientes a computadoras o robots "[…] a no ser que estuvieran hechos de las proteínas que constituyen los sistemas nerviosos", añadiendo a renglón seguido que "[…] en este caso no ofrecerían ningún enigma". Es más: en la estela de Ryle (1949), el propio Feigl descartaba de antemano todo intento de fijar el sustrato neurológico de las actitudes proposicionales, pues consideraba un error categorial tratar como un problema de articulación entre psicología y neurología lo que sería de hecho un problema de articulación entre psicología y lógica. Obviamente, invertir con cierta laxitud esta tesis daría como resultado la idea de que la explicación de la experiencia directa característica de las sensaciones crudas –a los ojos de Feigl (1958: 419, *supra*) el otro criterio decisivo de lo mental, junto con la intencionalidad y su manifestación en la inteligencia– resulta más asequible desde una óptica neurofisiológica.

Algo parecido ocurre en un trabajo que llegó a ser casi tan influyente como el de Feigl: *El redescubrimiento de la mente*, de John R. Searle (1992). En lo que atañe a las posibilidades del funcionalismo de rendir cuentas del carácter cualitativo, fenoménico de ciertos estados mentales, el diagnóstico de Searle es tan tajante como desesperanzador:

> The commonsense objection was that the functionalist seems to leave out the qualitative subjective feel of at least some of our mental states. There are certain quite specific qualitative experiences involved in seeing a red object or having a pain in the back, and just describing these experiences in terms of their causal relations leaves out these special *qualia*. (Searle 1992: 42)

Entre las razones que sustentan el dictamen destaca el hecho de que podemos concebir que cierto patrón de relaciones causales entre estímulos, respuestas y estados internos se diera sin ir acompañado del *quale* que nos interesa, o también viniendo acompañado de un carácter fenomenológico vivamente distinto –como en un *espectro invertido* (*cf. supra*). Ese hecho basta, a juicio de Searle, para desacreditar al funcionalismo –o al menos, con sus provocativas palabras, al "funcionalismo de caja negra":

> Now if this possibility is even intelligible to us—and it surely is—then black box functionalism must be wrong in supposing that neutrally specified causal relations are sufficient to account for mental phenomena; for such specifications leave out a crucial feature of many mental phenomena, namely, their qualitative feel. (Searle 1992: 43)

Respecto del fisicalismo, la posición de Searle es más confusa. Los materialistas, a su entender, rechazan la existencia de

> […] any irreducible phenomenological properties, such as consciousness or *qualia*. […] Why don't they just concede that these properties are ordinary, higher-order biological properties of neurophysiological systems such as human brains? (Searle 1992: 27-28)

La respuesta a esa pregunta –dice Searle a renglón seguido– "[…] es extremadamente compleja". Pero buena parte de la complejidad parece superflua si reparamos en que lo que Searle reclama a los materialistas que concedan es precisamente lo que cualquier materialista cabal defendería: que la consciencia, o los *qualia*, "[…] son propiedades biológicas ordinarias, de orden superior, de sistemas neurofisiológicos". Afirmar eso es equivalente a negar que sean "propiedades fenomenológicas irreductibles": no son irreductibles exactamente en la medida en que son propiedades biológicas ordinarias, de orden superior, de sistemas neurofisiológicos. Dado que Searle se afana en entender por qué los materialistas rechazan lo que de hecho conceden, no es raro que entenderlo se le haga extremadamente complejo.

El fondo de la cuestión es que Searle pretende mantener tanto que ciertas propiedades mentales son irreductibles a propiedades biológicas como que son propiedades biológicas: dicho de otro modo, que, pese a ser de hecho propiedades biológicas, son irreductibles a propiedades biológicas porque su existencia se enmarca intrínsecamente en una ontología de primera persona, y la de las demás propiedades biológicas –o conductuales, o funcionales– es una ontología de tercera persona.

> First and more important, there is the fact that you are now having certain unpleasant conscious sensations, and you are experiencing these sensations from your subjective, first-person point of view. It is these sensations that are constitutive of your present pain. But the pain is also caused by certain underlying neurophysiological processes consisting in large part of patterns of neuron firing in your thalamus and other regions of your brain. Now suppose we tried to reduce the subjective, conscious, first-person sensation of pain to the objective, third-person patterns of neuron firings. Suppose we tried to say the pain is really "nothing but" the patterns of neuron firings. Well, if we tried such an ontological reduction, the essential features of the pain would be left out. No description of the third-person, objective physiological facts would convey the subjective, first-person character of the pain, simply because the first-person features are different from the third-person features. (Searle 1992: 117)

Esas propiedades de primera persona que son irremediablemente diferentes de cualquier propiedad de tercera persona son, sin embargo, "propiedades biológicas ordinarias, de orden superior" y, por tanto, de tercera persona. La contradicción es flagrante, pero a lo ahora conviene atender no es a eso, sino al hecho de que este mismo trámite no se aplique cuando de lo que hablamos es de propiedades funcionales: bien podríamos decir, como el propio Searle, que ciertas propiedades fenomenológicas son propiedades funcionales y que, pese a ello, son irreductibles a propiedades funcionales a tenor de la ontología de primera persona que les es propia, y que no cuadra con la ontología de tercera persona a la que pertenecen las demás propiedades funcionales[160]. Quizá el fracaso de todo intento de entender la

[160] De hecho, es significativo que bajo cierta interpretación de las relaciones entre funcionalismo y fisicalismo –precisamente la que mantienen pensadores como D.M. Armstrong o J.C.C. Smart, a

consciencia y la subjetividad acabe por conducirnos a asumir resignadamente un esquema así –en el que ciertos fenómenos psicológicos se consideran simplemente fenómenos biológicos a pesar de que albergan determinadas propiedades que se diferencian radicalmente de cualquier otra propiedad biológica, y renunciamos a comprender cómo o por qué–, pero desde luego no parece un desiderátum particularmente pujante de cara a azuzar la investigación. Sea como sea, suponiendo que la maniobra de de Searle fuera legítima y propicia para la tesis de que las propiedades mentales son propiedades biológicas, no hay motivo para que no lo fuera también, como Searle parece pretender, para la de que las propiedades mentales son propiedades funcionales.

La tendencia, que Searle (1992) comparte con Feigl (1958), a ver en la identificación de una propiedad fenomenológica con una propiedad neurofisiológica algo que elucida *per se* la naturaleza de la primera en mayor medida que su identificación con una propiedad funcional se hace más transparente si cabe en la respuesta que da Searle a los argumentos de Thomas Nagel (1974, 1986). A lo que Searle opone resistencia es al planteamiento, que pone en boca de Nagel, según el cual:

> No possible account of neuronal behavior would explain why, given that behavior, we have to be, for example, in pain. No account could explain why pain was a necessary consequence of certain sorts of neuron firings. (Searle 1992: 101)

Debemos –dice Searle– rechazar tales conclusiones. Si el hecho de que el dolor sea una consecuencia necesaria de ciertos fenómenos neurofisiológicos nos parece "misterioso" es sólo debido a nuestra ignorancia, a que no tenemos más que una vaga idea de "[…] cómo funciona el sistema neurofisiología / consciencia", pero "[…] un conocimiento adecuado de su funcionamiento eliminaría el misterio" (Searle 1992: 102). Incluso el hecho de que podamos concebir cualquier fenómeno neurofisiológico cuya identidad con un fenómeno fenomenológico –digamos, el dolor– se nos plantee –es decir, que podamos concebir dicho fenómeno neurofisiológico sin que venga acompañado de dolor" se le antoja ahora a Searle un irrelevante fruto de nuestra ignorancia:

> Furthermore, the claim that we could always conceive of the possibility that certain brain states may not cause the appropriate conscious states might simply depend on our ignorance of how the brain works. Given a full understanding of the brain, it seems to me likely that we would think it obvious that if the brain was in a certain sort of state, it had to be conscious. (Searle 1992: 102)

---

quienes Searle (1992: 27) se refiere como "materialistas" que no aceptan que las propiedades fenomenológicas son propiedades biológicas ordinarias, de orden superior–, decir que las propiedades fenomenológicas son propiedades funcionales es justo lo mismo que decir que son propiedades biológicas de orden superior, de modo que el diferente rasero que Searle aplica en un caso y en otro se torna si cabe más palpable.

Pero no deja de resultar llamativo que ése fuera precisamente el argumento cardinal en el rechazo del funcionalismo por parte de Searle: que podamos imaginar cualquier patrón de relaciones funcionales en ausencia del fenómeno fenomenológico con el que presuntamente se identifican, o en presencia de otro distinto –que "[…] esa posibilidad sea siquiera concebible para nosotros" (Searle 1992: 43, *supra*)– era motivo más que suficiente para dar el funcionalismo por perdido. Cuando exactamente el mismo razonamiento se aplica a una concepción fisicalista de lo mental –que es lo que en este punto está en tela de juicio, aunque con la otra mano Searle parezca estar a la vez defendiendo una posición dualista–, sin embargo, todo cambia: inimaginables conocimientos futuros disiparán la ilusión de que no hay lazos necesarios entre propiedades biológicas y propiedades fenomenológicas[161], y la tesis de que la consciencia es una propiedad biológica quedará consagrada. Entretanto, según parece, mejor haríamos en conformarnos con aludir a ciertos "poderes causales" del cerebro, acaso en la estela de J.O. de la Mettrie y su abierta reivindicación de los poderes ocultos de la materia que poblaban la filosofía natural renacentista y que Descartes tanto empeño había puesto en impugnar, y aguardar así a que tan huidiza *vis sensitiva* sea por fin apresada sin mucha tardanza[162]. No: más

[161] O tal vez, después de todo, hallar tales lazos necesarios no sea, a su vez, necesario para dar por buena una explicación científica, puesto que "[…] not all explanations in science have the kind of necessity that we found in the relation between molecule movement and liquidity" (Searle 1992: 101). *Nihil obstat*, pero, de ser así, esa mayor generosidad en la concesión del rango de explicación satisfactoria debería, una vez más, otorgársele también a las propuestas funcionalistas que Searle rechaza taxativamente.

[162] *Nota bene*: "Considero el pensamiento tan poco incompatible con la materia organizada, que parece ser una propiedad de ésta, tal como la electricidad, la facultad motriz, la impenetrabilidad, la extensión, etc." (de la Mettrie 1748: 246). Que el materialismo que encardina el pensamiento de Searle –en la acaso inestable medida en que éste es materialista– comparte con el articulado por el autor de *El hombre máquina* cuando menos esa inmoderada confianza en unas propiedades de la materia aún desconocidas se refleja en ciertas afinidades con determinados estilemas del materialismo más –cabría decir– grosero. Así, por ejemplo, las polémicas afirmaciones de Karl Vogt (1847) en las que aseguraba que el pensamiento no es sino una secreción del cerebro, como lo es la bilis del hígado o la orina de los riñones –que casi al pie de la letra repetiría Carl Ludwig en su *Lehrbuch der Physiologie* (1852-1856, III: 345), donde sin duda las leerían tanto Iván Mijáilovich Séchenov como su joven discípulo Iván Petróvich Pávlov–, parecen tener un eco en la insistencia de Searle en que la consciencia es causada por procesos neurobiológicos "[…] and is as much a part of the natural biological order as […] photosynthesis, digestion, or mitosis" (Searle 1992: 90). La metáfora de la digestión, sea como sea, ya había sido profusamente empleada por Pierre Jean George Cabanis en sus *Rapports du physique et du moral de l'homme* (1802):

> Pour se faire une idée juste des opérations de la pensée, il faut considérer le cerveau comme un organe particulier destiné spécialement à la produire, de même que l'estomac et les intestins à opérer la digestion, le foie à filtrer la bile, les parotides et les glandes maxillaires et sublinguales à préparer les sucs salivaires. Les impressions, en arrivant au cerveau, le font entrer en activité, comme les aliments, en tombant dans l'estomac, l'excitent à la sécrétion plus abondante du sac gastrique et aux mouvements qui favorisent leur dissolution. La fonction propre de l'un est de se faire des images de chaque impression particulière, d'y attacher des signes, de combiner les différentes impressions, de les comparer entre elles, d'en tirer des jugements et des déterminations;

bien se diría que, como acertadamente concluye Pujadas (2002: 73) a la luz de consideraciones diferentes, "[…e]l problema de la consciencia sigue acechando sobre todas las teorías de la mente", y no es menos fiero ante unas que ante otras. En cualquier caso, Block y Fodor (1972a: 55) se aprestan a soslayar la cuestión de si el concepto de carácter cualitativo proscribe cualquier estirpe del funcionalismo –o incluso, como se ha apuntado, de sus ancestros filosóficos– por motivos estrictamente prácticos: de ser así, la validez del funcionalismo de tabla de máquina dejaría de ser un asunto que valiera la pena investigar.

Un parapeto del funcionalismo de tabla de máquina ensayado por Block y Fodor ante el asedio de los problemas ligados (*iii*) al carácter cualitativo de los estados mentales puede servir también de bisagra en relación con (*iv*) las dificultades del funcionalismo para cribar las diferencias *relevantes* entre estados mentales. Lo que se nos advierte es que:

> […] the proponent of *FSIT* […] might say that, given two functionally identical psychological states, we would (or perhaps "should") *take* them to be type identical, independent of their qualitative properties: that is, that differences between the qualitative properties of psychological states that do not determine corresponding functional differences are *ipso facto* irrelevant to the goals of theory construction in psychology, and hence should be ignored for purposes of type identification. (Block y Fodor 1972a: 54)

Convengamos, tentativamente, en que las diferencias cualitativas entre estados mentales que, contra las sospechas de Graham (2007: §5, *supra*), no originen diferencias funcionales son, precisamente por ello, superfluas a efectos de la tipología de estados mentales empleada en la explicación psicológica. Esto nos compromete, claro, con la premisa de que si *no* hay diferencias funcionales entre dos estados mentales, entonces ambos son, a efectos de la explicación psicológica, estados mentales del *mismo* tipo. La cuestión es ahora si habremos de respaldar también la premisa recíproca que asevera que cuando *sí* hay diferencias funcionales entre dos estados mentales, entonces ambos son, a efectos de la explicación psicológica, estados

---

> comme la fonction de l'autre est d'agir sur les substances nutritives dont la présence le stimule, de les dissoudre, d'en assimiler les sucs à notre nature.

Es claro, sin embargo, que ni el médico girondino ni Searle articulan la analogía de modo que el pensamiento resulte ser una secreción del cerebro, como hace Vogt, sino más bien su función: una retórica rudamente materialista, así pues, esconde una tesis no tan distante de un elemental funcionalismo. El rudo materialismo de Vogt, por otra parte, es desautorizado sin contemplaciones, entre otros, por Grey Walter (1957: 8), que omitiendo cortésmente el nombre del biólogo alemán lo tilda de "ridículo", y lo hace –lo cual no deja de resultar significativo– en un artículo escrito mientras era presidente de la Federación Internacional de Asociaciones de Electroencefalografía.

La metáfora digestiva, dicho sea de paso, puede rastrearse hasta un pensador tan poco sospechoso de incurrir en ese materialismo que se ha dado en llamar "vulgar" como el propio Descartes, quien asegura en las primeras páginas del *Discurso del Método* que "[…l]os que tienen más robusto razonar y digieren mejor sus pensamientos, para hacerlos claros e inteligibles, son los más capaces de llevar a los ánimos la persuasión" (1637: 72).

mentales de *distinto* tipo. Por supuesto, sería falaz suponer que la defensa del funcionalismo de tabla de máquina esbozada por Block y Fodor nos obligue a ello: de $\neg p \rightarrow q$ no se sigue $p \rightarrow \neg q$. El problema, más bien, es que aceptar esa premisa nos aboca a consecuencias desapacibles, pero el funcionalismo de tabla de máquina carece de recursos para rechazarla. Si cualquier diferencia funcional, por pequeña que sea, entre dos estados mentales los convierte en estados mentales de distinto tipo, la cosecha explicativa –al menos en el ámbito de la psicología humana– será parca. Bastará, por ejemplo, con que, entre dos personas que creen que seguirá lloviznando toda la tarde, una suela acompañar esa creencia de una vaga pesadumbre y la otra de una grata nostalgia, para que nos quede vetado considerar que ambos comparten la misma creencia –mejor dicho, que abrigan sendas creencias del mismo tipo–; tampoco podremos, por tanto, emplear esa coincidencia para explicar o predecir otras que puedan darse entre ellos[163]. Pero el funcionalismo de tabla de máquina gira precisamente en torno a la tesis de que dos estados mentales pertenecen al mismo tipo de estado mental si pertenecen al mismo tipo de estado de tabla de máquina (es decir, si son funcionalmente idénticos), y no en caso contrario.

Esta forma de enunciar uno de los preceptos capitales del funcionalismo de tabla de máquina ilustra bien el trasfondo del problema. Se proclama la identidad entre tipos de estados mentales y tipos de estados de autómatas abstractos; los estados mentales de un organismo se identifican con los estados de tabla de máquina postulados en la descripción óptima del organismo como autómata probabilista. Ahora bien, ¿en qué condiciones podríamos decir que dos estados de tabla de máquina son del mismo tipo –o, dado que hablamos de autómatas abstractos, que son en realidad el mismo estado de tabla de máquina? Una respuesta plausible a esa pregunta –en realidad, la más inmediata– es que la equivalencia de estados de tabla de máquina exige la equivalencia de tablas de máquina: el estado S_1 del autómata A y el estado S_2 del autómata B sólo pueden ser equivalentes si la totalidad de la tabla de máquina del autómata A y la totalidad de la tabla de máquina del autómata B son equivalentes –o sea, tratándose de autómatas abstractos que no podemos discernir por su instanciación física, si A y B son el mismo autómata. Pero esta respuesta, sumada al precepto de partida según el cual dos estados mentales son del mismo tipo si pertenecen al mismo tipo de estado de tabla de máquina, arroja la conclusión de que sólo dos organismos cuyas descripciones totales como autómatas probabilistas fueran equivalentes podrían llegar a atravesar sendos estados psicológicos del mismo tipo. Urge pues, acotar el requisito de identidad entre estados de tabla de máquina, pero la tarea se perfila inextricable. En una primera acometida intuitiva procuraríamos seguramente excluir de nuestro requisito de equivalencia con el estado S_1 aquellos fragmentos de tabla de máquina en los que no aparezcan referencias a ninguna de las aferencias, estados internos o eferencias que caractericen

[163] El ejemplo aducido por Block y Fodor hizo fortuna: dos personas no padecen el mismo tipo de dolor, según el funcionalismo de tabla de máquina, cuando se golpean el pulgar del pie si una de ellas suele reaccionar diciendo "Damn!" y otra diciendo "Darn!" –dos interjecciones muy frecuentes en inglés.

a S_1, ni a ninguna de las aferencias, estados internos o eferencias que caractericen a estados internos que aparezcan en la caracterización de S_1, etc., confiando a una sencilla formulación recursiva el resto de la labor de depuración. Una vez cumplimentado este trámite, podríamos establecer que S_2 pertenece al mismo tipo de estado de tabla de máquina que S_1 si y sólo si S_1 y S_2 tienen la misma caracterización funcional y los fragmentos de las tablas de máquina de A y B relevantes para la caracterización funcional de S_1 y S_2 son equivalentes. La estrategia parece conceptualmente impecable, pero nada garantiza que, para dos tablas de máquina dadas, permita aligerar significativamente, a efectos prácticos, el requisito de equivalencia entre estados de una y estados de otra. Es obviamente una cuestión empírica hasta qué punto estén entrelazadas las caracterizaciones funcionales de los estados de tabla de máquina de cada uno de los dos autómatas, así que es también una cuestión empírica de cuántos tramos de sus respectivas tablas de máquina se podría prescindir para fijar la equivalencia entre un estado de tabla de máquina de uno de ellos, y un estado de tabla de máquina del otro. En el caso de la psicología humana –para bien o para mal– lo mucho o lo poco que al respecto sabemos indica con rotundidad que el grado de entrecruzamiento de nuestros estados mentales es extremo: de hecho, resulta difícil concebir siquiera dos estados mentales que *no* pudieran estar ligados entre sí por medio de un número razonable de eslabones. Como queda claro en una constatación que Block y Fodor (1972a: 55) tildan de "embarazosa" para el funcionalismo de tabla de máquina:

> Indeed, on the assumption that there is a computational path from every state to every other, any two automata which have less than all their states in common will have none of their states in common. (Block y Fodor 1972a: 56)

El problema, como sagazmente aprecian Block y Fodor (1972a: 55), había sido ya barruntado por el propio Putnam (1960: 43), quien anticipaba que la principal dificultad de su modelo habría de radicar en "[…] pasar de modelos de organismos *específicos* a una forma *normal* de descripción de *organismos*"[164]. Sin embargo, no parece que los aprietos ocasionados por el holismo sean peculiares del funcionalismo de tabla de máquina –no en vano traslucen aquí, a poco que se piense en ello, las raíces holistas del problema del retorno de lo mental que ya aquejó el conductismo lógico[165].

[164] En su influyente ensayo sobre "La naturaleza de los estados mentales", Putnam (1967a: 299, *supra*), sin embargo, recuperaría el giro "forma normal" para referirse a un formalismo en el que pudieran expresarse leyes psicológicas supraespecíficas, un proyecto que se perfilaba como más practicable que el de hallar una forma normal para la teorización neurofisiológica sobre estados mentales.

[165] Aunque no se aborda en este estudio, la cuestión del holismo guarda estrechos lazos con las polémicas acerca de la caracterización del contenido semántico de los estados mentales en las que se ahondará infra. Una disección particularmente pulcra de dichos lazos puede encontrarse en Blanco (2000), en el marco de un elegante análisis de la noción de propiedad relacional.

La huella inequívoca del lenguaje es visible en las dos últimas brechas que según Block y Fodor (1972a) delata el escrutinio de la armadura del funcionalismo de tabla de máquina. Que nuestros estados mentales ostenten (*v*) una *productividad* indómita, que no se somete a límites de principio es, al igual que el hecho de que en ellos se desvele (*vi*) una disciplinada *sistematicidad*, fruto a todas luces del carácter proposicional que les confiere la imbricación del lenguaje en su misma estructura. Así, ante estados mentales netamente proposicionales –la creencia de *que* seguirá lloviendo toda la tarde, o de *que* la lluvia, en su debido tiempo y cuantía, es buena para las cosechas– tenemos pronta una retahíla de otros estados mentales posibles para el sujeto que los albergue –la creencia de que *no* seguirá lloviendo toda la tarde, la de que seguirá lloviendo toda la tarde *y toda la noche*, la de que seguirá lloviendo toda la tarde *y* la lluvia es, en su debido tiempo y cuantía, buena para las cosechas, etc.–, retahíla que cabe desplegar sin dato alguno acerca de la psicología del sujeto, apoyándonos únicamente en el conocimiento del lenguaje. Sin embargo, en un estado mental exento de tales deudas proposicionales –como la melancolía, fuera vaga pesadumbre o grata nostalgia, que envolvía al sujeto del ejemplo– la proliferación de estados mentales posibles se estanca: qué otros estados mentales pueda cobijar un sujeto del que sólo sabemos que siente una vaga pesadumbre es, en principio, estrictamente una cuestión empírica[166].

Pues bien, es precisamente ese notorio isomorfismo entre la estructura productiva y sistemática del conjunto de actitudes proposicionales posibles y la estructura del lenguaje lo que conduciría a Fodor (1975, *infra*) a concluir que la adecuada concepción de la naturaleza funcional de la mente exige la postulación de un código interno, productivo y sistemático, que provea de esos rasgos a nuestras actitudes proposicionales: es decir, un lenguaje del pensamiento. El funcionalismo de tabla de máquina se perfilaba como un planteamiento profundamente mal armado para estas lides, dado que identificaba el conjunto potencialmente infinito de estados mentales de un organismo con un conjunto, finito por definición, de estados de tabla de máquina, y dado además que carecía por completo de aparejos para dar cuenta de relaciones sistemáticas entre estados –aparte de listarlas. El hechizo de la idea de una *lingua mentis* se adivinaba ya en las conclusiones de Block y Fodor (1972a):

> Our point is that the same considerations [that apply to sentences] apply to the set of psychological states of an organism. Almost certainly, they too are, or at least include, a generated set, and their structural similarities correspond, at least in part, to similarities

[166] A fin de distinguir claramente entre la productividad y la sistematicidad de nuestras actitudes proposicionales, Block y Fodor aducen ejemplos que enmascaran los inextricables lazos entre ambas facetas del legado del lenguaje sobre los estados mentales que éste impregna. Aunque "*x* cree que 1+1=2", "*x* cree que 2+2=4", etc. sean un ejemplo prístino de productividad, y "*x* cree que *P*", "*x* cree que *P* y *Q*" lo sean de sistematicidad, si inspeccionamos creencias, deseos, temores o esperanzas verosímiles y sus posibles derivados, encontraremos por regla general –como en el ejemplo de la lluvia– una espesa mixtura en la que productividad y sistematicidad no son fácilmente segregables. Que productividad y sistematicidad son "dos aspectos del mismo fenómeno" es –en cualquier caso– expresamente recalcado por Block y Fodor (1972: 58, *infra*).

> in their derivation; that is, with psychological states as with sentences, the fact that they are productive and the fact that they exhibit internal structure are two aspects of the same phenomenon. If this is true, then a theory which fails to capture the structural relations within and among psychological states is overwhelmingly unlikely to arrive at a description adequate for the purposes of theoretical psychology. (Block y Fodor 1972a: 58)

En suma, los desperfectos que Block y Fodor (1972a) habían detectado en el mecanismo de clasificación de estados mentales en tipos del funcionalismo de tabla de máquina apelaban a hechos aparentemente cotidianos, como que en nuestra concepción de la mente cohabiten disposiciones, como la timidez, y estados actuales, como una vaga pesadumbre, que podamos vernos embargados por esa vaga pesadumbre a la vez que pensamos que seguirá lloviendo toda la tarde, que la coincidencia de ambas circunstancias pueda hacernos tararear una canción, que podamos compartir la creencia de que seguirá lloviendo toda la tarde con otra persona a quien, sin embargo, le evoca una grata nostalgia, que pudiéramos diferenciar la pesadumbre de la nostalgia incluso si las despertaran en nosotros las mismas circunstancias, y nos llevaran a idénticos pensamientos y costumbres, que esté en nosotros el poder de amparar, en principio, infinitas creencias o deseos, o que entre creencias o entre deseos haya lazos tan estrechos como entre la creencia de que seguirá lloviendo toda la tarde y la creencia de que no lo hará.

Algunos de esos desperfectos, una vez acotados los daños, merman por igual la credibilidad del funcionalismo de tabla de máquina que la de sus rivales –como se ha analizado minuciosamente. Así, los ahogos que el funcionalismo de tabla de máquina pueda sufrir para dar cuenta de (*i*) la distinción entre estados mentales disposicionales y actuales se reducen a la necesidad de esclarecer el papel de la consciencia en tal noción de actualización, tarea que al menos en el ámbito de las sensaciones y los estados emocionales queda entretejida con la de labrar una explicación de su (*iii*) carácter cualitativo. Pero, al igual que en lo que atañe a (*iv*) la carencia de criterios para descartar diferencias funcionales irrelevantes de cara a la construcción de una taxonomía de los estados mentales, sería injusto cargar las tintas contra el funcionalismo de tabla de máquina por sus déficits bajo estos conceptos cuando no hay indicios de que otras concepciones de la naturaleza de los estados mentales arrojen un balance más favorable al respecto[167]. En cambio, otros estragos resultaban más preocupantes: la incapacidad del funcionalismo de tabla de máquina para asimilar la concurrencia de (*ii*) estados mentales simultáneos y su papel conjunto en tanto que causas de una conducta o de otro estado mental, así como para rendir cuentas de la (*vi*) regimentación sistemática que permite que nuestras actitudes proposicionales exhiban (*v*) una estructura productiva, susceptible de desplegarse en un conjunto potencialmente infinito, eran descalabros sin paliativos.

[167] *Cf*., particularmente, los argumentos de Shapiro (2000, 2004, *infra*) contra la interpretación del funcionalismo como una tesis antirreduccionista, que dependen en un sentido crucial de una noción de diferencias fisiológicas relevantes que se revela parasitaria de criterios funcionales.

El título del trabajo de Block y Fodor (1972a) –"Lo que los estados psicológicos no son"– se revelaba al cabo como una diligente descripción de su contenido: ni el conductismo lógico, ni el fisicalismo de tipos, ni el prometedor funcionalismo de tabla de máquina lograban abastecernos de una concepción naturalista creíble de los estados mentales. Casi todo el trabajo estaba por hacer, aunque el horizonte parecía despejado:

> […I]f we wish to think of the psychology of organisms as represented by automata, then the psychological states of organisms seem to be analogous to the computational states of an automaton rather than to its machine table states. (Block y Fodor 1972a: 58)

El concepto de computación se convertiría así en la piedra angular del funcionalismo, como se ha anticipado de la mano de Rivière (1991b: 129, *supra*). De su mano, las nociones de regla y representación irían permeando la concepción funcionalista tanto de la explicación psicológica como de lo mental. Por computación no hemos de entender sino el proceso de manipulación automática de representaciones –símbolos discretos– con acuerdo a reglas que distinguen dichas representaciones según características formales –"sintácticas", se impondrá decir. Construir una teoría psicológica equivale, pues, a detallar las reglas y las representaciones que intervienen en el control de la conducta de determinados organismos o sistemas.

Como ha enfatizado Bechtel (1988: 158), esto implica que el funcionalismo computacional renuncia a la noción conductista de equivalencia psicológica derivada del examen orquestado por Turing (1950a, *supra*) para decidir si procede conceder el calificativo de "inteligente" al comportamiento de una máquina, y se inclina en cambio por una noción más severa de equivalencia psicológica, que exige no sólo indistinguibilidad de respuestas, sino también de procesos internos[168]. Pues bien, esclarecer la noción de equivalencia de procesos internos puede entenderse como el núcleo del empeño teórico de Pylyshyn (1980, 1984), acaso, junto con Fodor, el pensador que llegara a adquirir un compromiso más firme con el funcionalismo computacional. La necesidad de tal noción de equivalencia de procesos internos es para Pylyshyn una consecuencia natural de la lectura literal, no metafórica, de la

[168] Dicho de otro modo, el funcionalismo computacional introduciría la distinción entre el proyecto tecnológico de la inteligencia artificial y el proyecto teórico de la simulación cognitiva. Pero es patente que tal dicotomía ignora el interés teórico, y no sólo tecnológico, que bien puede tener de por sí el análisis de una tarea que el sistema cognitivo humano ejecute con mayor o menor eficacia, planteado con independencia de los procesos que efectivamente se pongan en juego cuando lo ejecuta. Es precisamente ese enfoque el que muchos de los pioneros del cognitivismo tomaron como propio; Marr (1982) acuñó para describirlo el concepto de nivel computacional de análisis, que opuso al nivel algorítmico y al de implementación –sólo una teoría que integrara los tres niveles de análisis llegaría a proporcionarnos, a su juicio, una comprensión cabal del sistema cognitivo. Más acertada es la constatación de Bechtel (1988: 160) de que el funcionalismo computacional encierra un estrecho compromiso con la concepción de la inteligencia artificial que Searle (1980) bautizó como "Inteligencia Artificial Fuerte", aunque es discutible que otras variantes del funcionalismo puedan sustraerse a tal compromiso –no podrán hacerlo, al menos, con sólo declararse solemnemente exentas de él, como se ha apuntado en relación con la posición de Gardner (1985: 412, *supra*).

homología entre cognición y computación que él mismo propugna (*cf*. por ejemplo Pylyshyn 1984: *xiii*, *supra*), y cuyas raíces entre los pioneros de las construcción de autómatas hemos estudiado en cierto detalle:

> This much is clear: In order that a computer program be viewed as a literal model of cognition, the program must correspond to the process people actually perform at a sufficiently fine and theoretically motivated level of resolution. In fact, in providing such a model what we would be claiming is that the performances of the model and the organism were produced "in the same way" or that they were carrying out the *same* process. To conclude that the model and the organism are carrying out the same process, we must impose independent constraints on what counts as the same process. This requires a principled notion of "strong equivalence" of processes. Obviously, such a notion is more refined than behavioral equivalence or mimicry […] (Pylyshyn 1984: xv)

Es patente, en efecto, la ruptura con el criterio de equivalencia que sustentaba el juego de imitación imaginado por Turing. No es una propuesta lo que Pylyshyn parece consignar en este punto, sino el registro de un consenso que sería consustancial al cognitivismo en psicología –y que había comenzado a fraguarse, como ya hemos apuntado, en Fodor (1968):

> This "black box" equivalence is now considered a weak equivalence criterion. […I]n the case of cognitive psychology, explanatory adequacy depends on a stronger sense of equivalence, particularly on knowing the details of the process at a suitable level of abstraction. (Pylyshyn 1984: 55)

La debilidad del criterio de equivalencia propuesto por Turing, con todo, no es a ojos de Pylyshyn una mera veleidad conductista. Antes al contrario, las capitales aportaciones teóricas que debemos a Turing, articuladas en torno a la noción de máquina universal, dependen de ese criterio de equivalencia. Lo que convierte a una máquina idealizada, abstracta, como la de Turing en un concepto central de la teoría de la computación es, según hemos estudiado detenidamente, la demostración de que hay una máquina tal que es capaz de ejecutar cualquier procedimiento efectivo, de naturaleza algorítmica, y, además, que el funcionamiento de cualquiera de dichas máquinas puede ser codificado de tal modo que una de ellas –la máquina universal– quede dotada de la capacidad de remedarlas a todas. Dicho de otro modo: si una tarea puede formalizarse de manera efectiva, entonces existe una máquina de Turing que puede realizarla –o más lacónicamente, tal como suele condensarse la tesis de Turing-Church, "computable" equivale a "Turing-computable"–; si el funcionamiento de una máquina puede describirse en el formalismo de la máquina de Turing, entonces puede ser simulado por la máquina universal. Como bien nos recuerda Pylyshyn (1984: 50), también las máquinas abstractas imaginadas por Emil Post y el cálculo lambda de Alonzo Church constituían formalismos completos, es decir, capaces de generar mecánicamente "todas las secuencias de expresiones susceptibles de ser interpretadas como pruebas y, por tanto, todos los teoremas demostrables de la lógica"; fue precisamente la constatación de la *equivalencia* entre

dichos formalismos lo que, como sabemos, dio forma a la tesis de Turing-Church. En efecto, el trabajo de Turing hacía patente la idea de que existe una cierta relación de equivalencia entre el procedimiento computacional desplegado por la máquina universal y el desplegado por la máquina que está siendo simulada, por muy diferentes que a simple vista pueda resultar ambos –ídem, cabe añadir, entre los procedimientos desplegados bajo distintos formalismos completos equivalentes:

> Here, the sense of formal equivalence is that the inputs and outputs of the U[niversal] M[achine] can be decoded uniformly as the inputs and outputs of the machine being simulated. We say that the UM "computes the same function" as the target machine, where, by same function, we mean the same input-output pairs or the same extension of the function. That is, there is no implication that the UM performs the same steps as the target machine. (Pylyshyn 1984: 50)

Así pues, disponer de ese criterio débil de equivalencia entre procesos computacionales, restringido a la extensión de la función computada, se revela como un elemento imprescindible para construir la noción de computabilidad a la manera de Turing. Lo que acaso podría achacarse a una veleidad conductista u operacionalista –seguramente comprensible, en todo caso, en el entorno intelectual de 1950– es ensanchar ese criterio de equivalencia más allá del concepto de computación, tratando de que ilumine nuestra comprensión del concepto de cognición o, como hacía Turing en "Computing Machinery and Intelligence", del de inteligencia.

Visto desde una perspectiva sólo levemente distinta: incluso si la totalidad del funcionamiento del sistema cognitivo humano resultara ser computable, *ergo* Turing-computable, y lográsemos proporcionar su descripción completa en el formalismo de la máquina de Turing y, con ello, su simulación completa –bajo determinadas idealizaciones, por supuesto, relativas a la naturaleza de aferencias y eferencias–, esto no contaría como una explicación satisfactoria del funcionamiento de nuestro sistema cognitivo. La razón, si Pylyshyn está en lo cierto, tiene que ver con el propio concepto de explicación:

> Explaining behavior is a much more stringent task than generating behavior. Specifically, explanation entails capturing the underlying generalizations as perspicuously as possible and relating them to certain universal principles. […] Thus […] explanatory adequacy requires a much more fine-grain correspondence between a computational model and an organism that is implied by input-output, or "Turing" equivalence. (Pylyshyn 1984: 53-54)

Parece obvio que la clave residirá, entonces, en cuál ha de ser la naturaleza de esa correspondencia para que un determinado modelo computacional pueda plegarse al canon de lo que consideraríamos una explicación solvente de cierta conducta. Dicho queda que se trataría de una relación más exigente que la equivalencia descrita por Turing (1950a, *supra*), pero, por otro lado:

> […] it is equally clear that […] computers not only are made of quite different materials than brains but, through details of realizing particular operations […], differ from the way the brain works. The correspondence between computational models and cognitive processes appears to fall somewhere between these extremes. (Pylyshyn 1984: 89)

Entre el extremo de equivalencia de caja negra y el de la estricta correspondencia entre los mecanismos físicos que ejecutan cada operación concreta, que ya rechazaran, como vimos, Bent Russell (1913) o Ross (1938), queda, desde luego, un extenso terreno: el que separa entre sí, si se quiere, las versiones más toscas del reduccionismo de inspiración conductista y del de inspiración neurológica. Así,

> […] there is plenty of scope in the possible claims a theory might make about the level of correspondence between model and empirical domain or about the properties of the model that could be said to have "psychological reality." (Pylyshyn 1984: 89)

La propuesta de Pylyshyn, en la que cristalizan buena parte de las intuiciones fundacionales del cognitivismo, es que el lugar que en ese vasto territorio es propicio para erigir una noción de correspondencia que nos provea de copiosas cosechas de explicaciones psicólogicamente robustas viene dado por el concepto de algoritmo, entendido como clase supraordinada a la de programa[169]. Ahora bien, que dos sistemas implementen instancias del mismo algoritmo entraña que recurran al mismo "conjunto de operaciones básicas"[170]. Dicho conjunto, que debe según Pylyshyn venir definido por consideraciones empíricas y teóricas independendientes, constituirá la arquitectura funcional del sistema, que sustenta los

[169] Ya Miller, Galanter y Pribram (1960, *infra*) proponían diferenciar tres niveles de abstracción en la explicación de la conducta: flujo de energía, flujo de información y flujo de control. Mucho mayor eco acabaría teniendo la propuesta de Newell (1982) de distinguir en la descripción de sistemas de representación del conocimiento un *nivel simbólico* y un *nivel de conocimiento* –además de tres niveles no semánticos: un nivel de dispositivo, un nivel de circuito y un nivel lógico. El propio Pylyshyn (1984: 24) hace suya expresamente la distinción de Newell, si bien parece inclinarse por discernir entre un nivel físico o biológico, un nivel simbólico o sintáctico, y un nivel semántico o intencional (*cf.* por ejemplo Pylyshyn 1984: 259). También la exhortación de Marr (1982, supra) a diferenciar el nivel computacional de análisis del nivel algorítmico y el de implementación –y a integrar los tres en nuestro quehacer teórico– está, naturalmente, emparentada con éstas. Estudiar a fondo los distintos usos que la noción de *nivel* ha recibido en el seno del funcionalismo y el cognitivismo sería, indudablemente, una contribución de primer orden a la elucidación de la armazón teórica de esas concepciones de lo mental.

[170] Expresada con mayor precisión, la tesis de Pylyshyn es que dos programas:

> […] can be thought of as strongly equivalent or as different realizations of the same algorithm or the same cognitive process if they can be represented by the same program in some theoretically specified virtual machine. […] The formal structure of the virtual machine […] thus represents the theoretical definition of, for example, the right level of specificity […] at which to view mental processes, the sort of functional resources the brain makes available –what operations are primitive, how memory is organized and accessed, what sequences are allowed, what limitations exist on the passing of arguments and on the capacities of various buffers, and so on. (Pylyshyn 1984: 91-92)

procesos computacionales, de naturaleza representacional, que el sistema despliega sin requerir ella misma una explicación en tales términos:

> By "functional architecture" I mean those basic information-processing mechanisms of the system for which a nonrepresentational or nonsemantic account is sufficient. The operation of the functional architecture might be explained in physical or biological terms, or it might simply be characterized in functional terms when the relevant biological mechanisms are not known. (Pylyshyn 1984: *xv-xvi*)

Así pues, como quiera que un modelo computacional adecuado de un proceso cognitivo debe –si ha de ceñirse a los requisitos de adecuación explicativa de Pylyshyn– emplear el mismo algoritmo que empleen los sujetos, y como quiera que qué algoritmos pueda emplear depende en buena medida de la arquitectura funcional de la que se le dote, la elección de la arquitectura funcional idónea se perfila como una decisión capital en la construcción de modelos cognitivos, decisión que –insiste Pylyshsyn– debe descansar sobre una firme justificación empírica[171].

La concepción de la mente que se ha dado en conocer como *funcionalismo homuncular* tiene precisamente en William Bechtel (1988) uno de sus principales adalides[172]. No es raro, pues, que Bechtel describa el funcionalismo homuncular como un logrado intento de zanjar los apuros que afligen a otras variedades de funcionalismo empírico –en especial, al funcionalismo computacional predicado por Fodor, Pylyshyn o Block. A menudo pasa desapercibido, sin embargo, el propio hecho de que el funcionalismo homuncular se perfile como una alternativa al funcionalismo computacional, de que se arrogue la potestad de plantear un diálogo, por así decir, de igual a igual. Pero que el funcionalismo homuncular constituya una concepción de la mente sustancialmente distinta del funcionalismo computacional es, a su vez, una tesis sustancial, y procede examinar qué apoyo puede procurársele.

El mismo Bechtel (1988: 163) toma nota de una observación de Dennett (1975) –en uno de los trabajos seminales de la perspectiva homuncularista– respecto a que la estrategia de dividir la tarea cognitiva que pretendemos explicar en subtareas cada vez más simples, hasta alcanzar "[…] problemas o descripciones de tareas que son obviamente mecánicas" (Dennett 1975: 80), es *de hecho* puesta en práctica por la mayoría de los investigadores en inteligencia artificial –incluso, se sobreentiende,

---

[171] Recurrir a la noción de máquina virtual permite de nuevo una formulación más concisa y exacta: la arquitectura funcional compartida por dos sistemas que ejecutan programas que son instancias del mismo algoritmo es la estructura formal de la máquina virtual en la que ambos programas resultan ser el mismo. En suma,

> […] we individuate cognitive processes in terms of the expression in the canonical language of this virtual machine.[…] Specifying the functional architecture of a system is like providing a manual that defines some particular programming language. Indeed, defining a programming language is equivalent to specifying the functional architecture of a virtual machine. (Pylyshyn 1984: 91-92)

[172] Otro de sus más ilustres prelados, William G. Lycan, suele denominarla con la contracción "homuncionalismo"; acaso "homúnculo-funcionalismo" resulte menos abrupto a nuestros oídos

aunque declaren adherirse a alguna versión del funcionalismo computacional. Desde luego, la caracterización que Bechtel proporciona del funcionalismo homuncular, desplegada sobre la plantilla del modelo de explicación funcional que había confeccionado Cummins (1975, 1983), sería aplicable al pie de la letra al funcionalismo computacional. Con todo –arguye Bechtel:

> Aunque existe esta afinidad entre el Funcionalismo Homuncular y el Funcionalismo Computacional, el punto focal es diferente. La meta de la mayor parte de los investigadores de IA […] es sintética: diseñar un programa para realizar la tarea global. La estructura jerárquica de un programa no es, en última instancia, crítica, y las operaciones de las subrutinas son de una pieza respecto de las operaciones del programa principal. Sin embargo, para el Funcionalista Homuncular la estructura jerárquica resulta más importante. El Funcionalismo Homuncular trata los recuadros de las cartas de flujo como algo que caracteriza unidades modulares efectivas que efectúan sus propias actividades. (Bechtel 1988: 163)

La velada acusación según la cual el funcionalismo computacional no prestaría la debida atención a las jerarquías funcionales subyacentes a las competencias cognitivas que trata de explicar no es fácil de fundamentar. Uno de los motivos decisivos para que Block y Fodor abanderasen la moción contra el funcionalismo de tabla de máquina y en favor del funcionalismo computacional –recordemos– era, precisamente, que "[…] if probabilistic automata are to be used as models of an organism, the appropiate model will be a set of intercommunicating automata operating in parallel" (Block y Fodor 1972a: 53, *supra*). Incluso la terminología elegida por Bechtel para describir los elementos –"las unidades modulares efectivas"– de esa estructura jerárquica que, a su entender, es menospreciada por el funcionalismo computacional evoca vívidamente la concepción de la arquitectura de las facultades psicológicas propugnada, con la irónica vehemencia que es característica de sus trabajos, por el propio Fodor (1983, 1986b). Así que será preciso dar con alguna razón más rotunda para considerar el funcionalismo homuncular como una concepción de la mente significativamente distinta del funcionalismo computacional. No habrá que rebuscar mucho, pues Bechtel la esboza a renglón seguido:

> Lo mismo que al sistema global se le atribuyen creencias y deseos, los funcionalistas homunculares como Dennett también atribuyen creencias y deseos a los homúnculos que constituyen el sistema. Las creencias y deseos de un homúnculo serán diferentes de las que se atribuyen al sistema total: serán creencias y deseos sobre las tareas a realizar por el homúnculo […]. (Bechtel 1988: 163)

Pese a lo aparatosas que resultan desde este punto de vista las diferencias entre la posición de Dennett y la de –digamos– Fodor, el funcionalismo homuncular no vendría a constituir, de ser ésta la peculiaridad de su sentido, una alternativa al funcionalismo computacional tanto como una extensión de su alcance: una exhortación a ensanchar nuestra concepción de lo mental de modo tal que consintamos en atribuir actitudes proposicionales –o, mejor dicho, estados *análogos a*

las creencias y los deseos, conceptualizados según exija el desarrollo científico de la psicología– no sólo a ciertos organismos sino también a algunas de sus partes –a aquellas, claro está, que acataran determinadas caracterizaciones funcionales. Pero todo esto no es otra cosa que la conjunción de un enfoque nítidamente funcionalista de los estados mentales, apenas discernible del auspiciado por el funcionalismo computacional, y la concepción instrumentalista de la intencionalidad que viene desarrollando Dennett (1981, 1987), según la cual ésta, como suele decirse de la belleza, reside en el ojo del observador. El resultado, por lo demás, aunque sin duda radicaliza el sentido de la metáfora homuncular, lo hace a costa de incurrir en una doctrina sobre la naturaleza de los fenómenos psicológicos que es sumamente vulnerable –salvo que se trate de una estipulación léxica, en cuyo caso es sumamente cuestionable. La razón es sencilla: bien podría darse el caso de que el desarrollo científico de la psicología viniera precisamente a reconocer en las creencias y deseos atribuidas al sistema global determinadas propiedades que, ausentes en los estados funcionales de los subsistemas –homúnculos– postulados por la propia teoría, justificaran la atribución a aquellas, pero no a éstos, de contenido intencional, semántico, *en sentido análogo al coloquial*.

Con todo, no son pocas las intuiciones tenazmente perseguidas por los partidarios de la concepción homuncular de la mente en las que germinan contribuciones cabales a la comprensión de la naturaleza de lo mental que pueda proporcionarnos el funcionalismo. La insistencia –por ejemplo– de Lycan (1987: 38, *infra*) en suavizar la distinción entre lo estructural y lo funcional, rechazando que se trate de categorías estancas e insensibles al punto de vista desde el que se articule nuestro análisis, tiene la virtud de enlazar el aparato conceptual funcionalista con el del hilemorfismo, dando cuerpo a los indicios de una vigencia del aristotelismo en el funcionalismo más honda que la que ya señalara Pylyshyn (1984: 49) al apuntar que "[…] the structure-function or form-substance distinction implicit in notions of symbol and mechanisms dates from the ancient Greeks" –además, como veremos, la propuesta de Lycan logra articular en buena medida la réplica del funcionalismo, entendido como una posición antifisicalista, a las crecientes acometidas del reduccionismo de inspiración neurofisiológica.

También se engrana el funcionalismo con algunas ideas capitales del pensamiento aristotélico merced al énfasis en el trasfondo telelógico de la caracterización funcional de los estados mentales, que tanto Lycan (1987) como Bechtel (1988) cultivan, y que ya en Millikan (1984) había empezado a cobrar su expresión más acabada. Además, como apunta Bechtel (1988: 185), la incorporación de nociones teleológicas al análisis funcionalista logra entretejerlo con la tradición funcionalista que a través de los trabajos de James y Angell impregna desde sus inicios a la psicología científica[173]. Incluso en los planteamientos más acendradamente

[173] Chacón (2001: 117), por ejemplo, condensa acertadamente el ímpetu del funcionalismo decimonónico como el de "[…] un movimiento que […] reivindicó la necesidad de insertar el estudio científico de la conducta humana en su dimensión biológica, como conducta *adaptativa a* un medio y *orientada a* la satisfacción de sus necesidades"; así, a su entender, "[…e]l estudio de los actos humanos

operacionalistas de Skinner pervive, por lo demás, esa vocación de relectura de las causas finales, que una vez más entrelaza su pensamiento con la genésis y el desarrollo de un cognitivismo que él, acaso más que nadie, denostaba. En efecto, Skinner (1974: 54) supo ver la colosal fuerza de la idea de que "[…] la conducta operante […] es el campo del propósito y la intención" –también la inteligencia– de los organismos en el mismo sentido que la selección natural es el del propósito, la intención y el entendimiento de la divinidad creadora. En la atinada síntesis de Ringen (1990), la tesis cardinal de Skinner a este respecto es que:

> […] the causal processes producing the behavior traditionally called purposive and intentional are instances of selection by consequences, a causal mode exhibited in the analogous processes of operant conditioning (the contingencies of reinforcement) and natural selection (the contingencies of survival). (Ringen 1990: 168)

Acaso huelgue añadir ya que la descripción de los modos en que esa "selección por consecuencias" moldean la conducta de los organismos se volvería inconmensurablemente más intricada desde el momento en que se admitiera –entre las airadas protestas de Skinner (1985, 1987, 1989, 1990)– que abarcan también la determinación de la naturaleza y, en especial, del contenido intencional de estados internos de dichos organismos capaces de operar como causas de su conducta. El propio Ringen (1990: 175), por lo demás, anota la afinidad entre el énfasis de Skinner en los procesos de selección por consecuencias y las concepciones teleológicas de la semántica de los estados mentales, de las que toma como patrón el trabajo de Millikan (1984, 1993).

También Lycan (1987: 137) menciona a Millikan (1984) –al lado de Popper (1972), Wimsatt (1972), Wright (1973) y los trabajos entonces inéditos de K. Neander– como fuentes de inspiración de su relectura teleológica del funcionalismo. En consonancia con su insistencia en desdibujar los límites entre las nociones de estructura y función (Lycan 1987: 38, *infra*), la idea que Lycan trata de afianzar en el seno del debate sobre la naturaleza de lo teleológico es que el carácter teleológico de una determinada descripción de la realidad es una cuestión de grado. Así, en diferentes niveles de la naturaleza –condensaciones de regularidades nómicas, como los concibe Lycan– la caracterización de un mismo objeto o evento resultará *más o menos teleológica*. Vale la pena recordar el ingenioso ejemplo invocado por Lycan (1987: 43) y las conclusiones que lo siguen:

> One and the same space-time slice may be occupied by a collection of molecules, a piece of very hard stuff, a metal strip with an articulated flange, a mover of tumblers, a key, an unlocker of doors, an allower of entry to hotel rooms, a facilitator of adulterous liaisons, a

como *función de* un organismo vivo vendría […] a superar las limitaciones del análisis atomista de los contenidos de conciencia, tal como había sido planteado inicialmente por los primeros psicólogos experimentales". La vindicación de lo teleológico –si puede haber una vindicación en la radical reinterpretación de su naturaleza que alentaba el darwinismo–, es tan patente que las cursivas que aquí se añaden para recalcarla acaso resulten superfluas.

> destroyer of souls. Thus, we cannot split our theory of nature neatly into a well-behaved, purely mechanistic part and a dubious, messy vitalistic part better ignored or done away with.

Esa gradualidad, por supuesto, es cara a Lycan: sin ella, difícilmente podría sostener su convicción de que el homuncularismo es la heurística más provechosa para nuestra comprensión de lo mental, pues el abismo entre lo teleológico y lo mecánico desarmaría la supuesta homogeneidad de la sucesión de niveles explicativos en la que la postulación de homúnculos funcionales adquiriría legitimidad.

Sea como fuera, no conviene, de nuevo, dar por sentado que el funcionalismo teleológico constituya una ruptura drástica con el espíritu de las reflexiones de Putnam que dieron vida a la concepción funcionalista de lo mental, aunque sí con los pormenores de aquel funcionalismo de tabla de máquina. Ya el trabajo de Putnam (1967a) que abrió el camino de la concepción funcionalista de lo mental, la especificación del papel desempeñado en la organización funcional del organismo por las aferencias –o, se sobreentiende, eferencias– características de un determinado estado mental conllevaba una mención –si bien pasajera– de la función de los órganos responsables de dichas aferencias. En efecto, tras recalcar que el análisis disposicional del estado de sentir dolor no puede ir mucho más allá de describirlo como la disposición del organismo a comportarse como si sintiera dolor (Putnam 1967a: 229, *supra*), Putnam enfatiza el modo en que el funcionalismo puede imponerse a esa dificultad:

> In contrast, we *can* specify the functional state with which we propose to identify pain, at least roughly, without using the notion of pain. Namely, the functional state we have in mind is the state of receiving sensory inputs which play a certain role in the Functional Organization of the organism. This role is characterized, at least partially, by the fact that the sense organs responsible for the inputs in question are organs whose function is to detect damage to the body, or dangerous extremes of temperature, pressure, etc., and by the fact that the "input" themselves, whatever their physical realization, represent a condition that the organism assigns a high disvalue to. (Putnam 1967a: 229)

Por otra parte, la atemperación de la dicotomía de estructura y función ensayada por Lycan permite dilatar el contorno del funcionalismo hasta hacerlo cobijar en su seno incluso al propio fisicalismo, sometido, eso sí, a una profunda relectura. La teoría de la identidad psicofísica se entiende entonces como la tesis de que las propiedades de un organismo que se identifican con sus estados mentales quedarán adecuadamente descritas en el mismo rango de la gradación que va de lo estructural a lo funcional en el que tradicionalmente viene describiéndolas la fisiología: se trataría así, sencillamente, de una versión del funcionalismo que admite menos capas de caracterización funcional en la descripción de los estados mentales de las que acostumbran a recomendar los funcionalistas.

> […] si aceptamos también mi afirmación de que las caracterizaciones homunculares y las caracterizaciones fisiológicas de los estados de personas reflejan meramente diferentes

> niveles de abstracción dentro de una jerarquía o continuo funcional circundante entonces ya no podemos distinguir al funcionalista del teórico de la identidad de ninguna manera absoluta. 'Neurona', por ejemplo, puede entenderse *o* como un término fisiológico (que denota un género de célula humana) *o* como un término (teleo-)funcional (que denota un relé de carga eléctrica) […]. Así, pues, *incluso el teórico de la identidad es funcionalista*: alguien que coloca a las entidades mentales en un nivel muy bajo de abstracción. (Lycan 1981: 47 *apud* Bechtel 1988: 164)

Adviértase, sin embargo, que la metáfora homuncular es completamente inerte en la argumentación: que sea o no viable interpretar el fisicalismo como un funcionalismo particularmente reacio a la abstracción en nada depende de que sea aconsejable o no atribuir a las partes del sistema cognitivo distinguidas por nuestra caracterización funcional –como, *in extremis*, las neuronas, o los canales sinápticos– estados análogos a los que acostumbramos a atribuir coloquialmente a nuestros semejantes –o sea: creencias, deseos, temores. Eso, desde luego, es una buena noticia para Lycan, pues libera su análisis de todo compromiso con los aspectos más controvertidos de la concepción de lo mental encabezada por Dennett. No es ésa, sin embargo, la interpretación de los vínculos entre homuncularismo y funcionalismo teleológico que despliega Bechtel. Antes al contrario, a su entender:

> […] el Funcionalismo Teleológico es un complemento natural del Funcionalismo Homuncular. (Bechtel 1988: 185)

Pero el razonamiento que precede a esta conclusión es sumamente confuso:

> […E]l Funcionalismo Homuncular empieza con una explicación de lo que lleva a cabo todo el sistema y a continuación intenta explicar esa realización descomponiendo ese sistema en subsistemas (homúnculos). La perspectiva teleológica entra en escena con la manera en la que especificamos las tareas que el sistema está realizando. Si hacemos eso usando expresiones idiomáticas intencionales y si adoptamos una perspectiva evolucionista sobre la intencionalidad, ya hemos introducido una perspectiva teleológica. Estamos tratando los estados mentales como estados adaptativos de organismos. (Bechtel 1988: 184)

Son muchos los resquicios del argumento. Uno: si especificamos mediante giros intencionales las tareas realizadas por un sistema y mantenemos una visión evolucionista de la intencionalidad, habremos introducido una perspectiva teleológica sólo en la medida en que nuestra concepción de los procesos evolutivos sea también teleológica. Dos: en todo caso, lo que estará entonces del todo ausente de nuestro análisis será la perspectiva homuncular. Tres: si especificamos también mediante giros intencionales las tareas que están realizando los subsistemas funcionalmente definidos para el sistema estudiado, habremos, entonces sí, introducido –o, más bien, estipulado– una perspectiva homuncular, pero quedaremos *ipso facto* obligados a respaldarla con una perspectiva evolucionista sobre la actividad de dichos subsistemas tal que, dada la concepción evolucionista de

la intencionalidad que se nos supone, el empleo de los giros intencionales en cuestión quede justificado. Cuatro: este último requisito no se cumple únicamente con señalar que la actividad de los subsistemas cuya caracterización como homúnculos se encuentra en litigio viene moldeada por presiones evolutivas, puesto que lo mismo puede decirse de cualquier otro subsistema del organismo, desde los alveolos pulmonares hasta las glándulas sudoríparas; la réplica según la cual cualquier otro subsistema del organismo puede en efecto ser legítimamente caracterizado en términos homunculares incurre en clara *petitio principii*, aparte de que resulta si cabe más controvertida que la tesis que se discute. Cinco: proporcionar el fundamento adaptacionista que se reclama para la tesis homuncular exigiría explicitar analogías significativas entre la naturaleza de las presiones evolutivas que dotan de carácter intencional a los estados mentales del organismo y las que presuntamente hacen lo propio con los estados de los subsistemas descritos en jerga homuncular. En suma: ni defender que la intencionalidad ha de entenderse atendiendo a las presiones evolutivas a las que están o han estado sometidos los organismos que la albergan equivale a concebir la intencionalidad bajo cánones teleológicos, ni muchísimo menos equivale a defender que *cualquier sistema o subsistema sometido a cualesquiera presiones evolutivas* alberga por ello estados intencionales de la misma naturaleza que los que atribuimos a nuestros congéneres.

La prolongada controversia entre tan diversas interpretaciones del funcionalismo, que bien podría llenar muchas más páginas, parece dificultar la circunscripción de un territorio compartido por todas ellas, acaso porque, al fin y al cabo, tal territorio sea en verdad exiguo. No está de más, ante tales dificultades, recordar como un claro día del verano ático, mientras en su afán por entender la naturaleza del amor conversaban acerca de las distintas clases de delirio, Sócrates apuntaría para el joven e inocente Fedro la enseñanza de que "[…] hay que poder dividir las ideas siguiendo sus naturales articulaciones, y no ponerse a quebrantar ninguno de sus miembros a manera de un mal carnicero" (*Fedro* 265*e*) –la división aparecía así, junto con la definición, como la médula de la dialéctica. Fuera del convencimiento –heredado de Sócrates– de que ha de existir un esquema taxonómico que describa lo mental según las costuras de la realidad, fuera de la insistencia, fruto de ese convencimiento, en preguntarse cuáles son los criterios adecuados para clasificar los estados mentales, lo que la concepción funcionalista de la mente parece dar por acordado en todas sus vertientes no es más que la tesis mínima de que esa taxonomía certera habrá de apelar a las relaciones causales entabladas por los fenómenos que trate de clasificar. Como lacónicamente anota Bechtel, en definitiva:

> El Funcionalismo mantiene que los eventos mentales se clasifican en términos de sus papeles causales. (Bechtel 1988: 150)

Espíritu, materia, función. Lecturas ontológicas del funcionalismo

Una simplificada distinción entre ontología y metafísica –según la cual ésta es la parte de aquella que se ocupa de la taxonomía de lo que existe– sirve a Block (2007b) para ubicar conceptualmente al funcionalismo respecto del fisicalismo y el dualismo. La discrepancia entre dualistas y fisicalistas sería tanto ontológica como metafísica: sobre si existen o no sustancias o propiedades inmateriales, tanto como sobre si la naturaleza de los estados mentales –o cualquiera que sea el objeto de la disputa– es metafísicamente material o inmaterial, es decir, si se clasifican o no según podamos atribuirles tales o cuales propiedades o atribuírselos a tales o cuales sustancias. Pero el funcionalismo no tercia en la querella ontológica. Por el contrario, se limita a afirmar que la clasificación de estados mentales en tipos debe proceder según criterios funcionales, que lo que hace que una instancia de dolor sea dolor es algo de orden funcional: su tesis concierne sólo a la metafísica del dolor –de los fenómenos psicológicos en general. Así pues:

> An adding automaton is defined by functional relations to its inputs and outputs, but this does not rule out and adding automaton powered by an immaterial soul. (Block 2007b: 9)

O bien, una vez trasladada la conclusión al caso más disputado de la mente humana:

> Functionalism tells us what pains have in common –what makes them pains– is their function; but functionalism does not tell us whether the beings that have pains have any nonphysical parts. (Block 1996: 19)

Efectivamente, la compatibilidad de funcionalismo y dualismo es patente, como lo es la compatibilidad de funcionalismo y fisicalismo de casos. Por supuesto, aunque la distinción entre alcance de tipos y alcance de casos no suela aplicarse al dualismo, resulta claro que una concepción dualista de lo mental formulada bajo una óptica de tipos –donde, *grosso modo*, "Este pensamiento particular es un particular estado espiritual" diera paso a "Todo pensamiento de este tipo es un estado espiritual de este tipo"– colisionaría tan violentamente con el funcionalismo como pueda hacerlo el fisicalismo de tipos[174]. Ahora bien, una interpretación con alcance de tipos es sin duda, a falta de distinciones explícitas que exigirían un minucioso estudio histórico y filológico, la lectura natural de la tradición dualista, aunque sólo sea porque la modestia de las ambiciones explicativas es en ella tan poco característica como en la

[174] Analizar las relaciones entre fisicalismo, dualismo y funcionalismo recurriendo a la distinción entre alcance de tipos y alcance de casos muestra que plantear el asunto –como hace Block– de un modo que nos obliga a avanzar un compromiso en cuanto a las tareas propias de la metafísica y de la ontología suponía una complicación innecesaria: basta constatar que el funcionalismo puede tan bien conciliarse con el dualismo como con el materialismo si uno u otro se interpretan con alcance de casos, pero colisiona con ambos si estos se entienden con alcance de tipos.

tradición fisicalista. Así que la distante fraternidad entre funcionalismo y dualismo que se nos bosqueja descansa sobre un retrato del dualismo en que éste aparece grotescamente disminuido. Las desavenencias surgirían tan pronto el dualismo reclamara su propio vigor[175].

También ha reparado Block (1980a: 35, 1997: 20) en la peculiar circunstancia de que algunos destacados ponentes del funcionalismo –como Putnam o Fodor, *cf.* por ejemplo Fodor (1985: 15, *supra*)– consideren que su propuesta resulta incompatible con el fisicalismo, mientras otros –siguiendo a Armstrong (1968) y Lewis (1969)– lo han visto precisamente como una reivindicación del fisicalismo. La confrontación entre ambas posiciones lleva a Block a distinguir la "tesis de identidad de especificaciones funcionales", que defenderían Lewis y Armstrong, de la "tesis de identidad de estados funcionales", que habrían propugnado Fodor y Putnam. Lo que Armstrong o Lewis sostendrían es la identidad entre un tipo de estado mental y un tipo de estado físico especificado según criterios funcionales –hasta donde sabemos, un tipo de estado cerebral, pero esto podría no ser así en otros sistemas cognitivos.

---

[175] A pesar del afán con el que los adalides del funcionalismo han reivindicado su condición de materialistas (*cf.*, por ejemplo, Fodor 1974, *supra*), se ha hablado en ocasiones del funcionalismo como un nuevo dualismo –el "dualismo funcionalista", dice Rivière (1991b: 139). Existe cierta perspectiva, claro, desde la que todo lo que no sea el eliminacionismo es un dualismo: incluso la tesis de identidad psicofísica, tal como había sido formulada por Smart (1959), fue recusada por Stevenson (1960) y Bradley (1964) –así como, siquiera temporalmente, por el propio Smart (1961, 1967)–, quienes no veía en ellas sino una recaída en el dualismo de propiedades (*supra*). Es harto improbable, desde luego, que sea ésa la perspectiva en la que opera el pensamiento de Rivière.

Más razonable sería pensar que Rivière aluda al hecho de que el funcionalismo puede interpretarse como un dualismo epistemológico, en el sentido de que diferencia entre dos niveles de explicación autónomos –neurológico y psicológico–, o incluso concede alguna suerte de privilegio al segundo. Hemos pasado ya revista a la insistencia de Lycan (1987) o Bechtel y Mundale (1999), contra esa tendencia a la dicotomización, en la multiplicidad de niveles explicativos desde los que cabe articular nuestro escrutinio de la realidad; también a los argumentos de tintes pragmatistas de Hardcastle (1996) al respecto. En todo caso, no es difícil mostrar, como se ha esbozado *supra*, que esa multiplicidad de niveles estaba ya presente en las fuentes del cognitivismo: sólo hace falta una lectura atenta de Miller, Galanter y Pribram (1960), Newell (1982) o Pylyshyn (1984). En Bechtel (1994) puede encontrarse un interesantísimo análisis de la propia noción de nivel explicativo, y el uso que ha recibido en el seno de las ciencias cognitivas.

Conviene anotar, sin embargo, que en el mismo trabajo en el que clasifica al funcionalismo como una variedad de dualismo, Rivière (1991b: 151) subraya la afirmación contraria respecto del paradigma conexionista: que este constituye "[…] una alternativa realista, no dualista" de solución al "[…] viejísimo problema […] de las relaciones entre una sustanci extensional, el cuerpo, y un conjunto de funcionales intencionales, al que llamamos mente". Fundamentar la distinción en virtud de la cual el cognitivismo de inspiración algorítmica –*clásico*, como suele decirse– sería una concepción dualista de la relación entre mente y cuerpo y no lo sería en cambio el conexionismo requiere sin duda un esfuerzo mayor que el que le dedica Rivière, que se limita a señalar que en la aproximación conexionista "[…] lo que se computan no son 'símbolos' […] sino variables subsimbólicas de unidades moleculares de cómputo, que actúan en paralelo y con arreglo a leyes precisas" (Rivière 1991b: 151). Por cierta que sea tal cosa, difícilmente puede tomarse como una reconstrucción de la diferencia entre dualismo y "realismo", al menos en los términos del "[…] viejísimo problema" al que alude Rivière.

Esto es a su parecer todo lo que hace falta para que las tesis funcionalistas cuenten como tesis fisicalistas con alcance de tipos. Tal como describiendo la posición de Lewis se ocupa de recalcar Bechtel (1988: 150), "[…] la identificación funcional de los estados mentales […] proporciona la base para una identificación subsiguiente de qué estados físicos los instancian". En este sentido, la evocación de los orígenes del funcionalismo en la memoria de Lewis resulta muy distinta de la que trasluce en los recuerdos de Fodor (1985: 15, *supra*):

> A dozen years or so ago, D.M. Armstrong and I (independently) proposed a materialist theory of mind that joins claims of type-type psychophysical identity with a behaviorist or functionalist way of characterizing mental states such as pain. (Lewis 1980: 217)

Por otro lado, ya Lewis (1969) se había exonerado de la defensa de un fisicalismo ajeno a los planteamientos funcionalistas que sólo residiría en la imaginación del crítico. Así que tratar de combatir la concepción del funcionalismo como una tesis de identidad psicofísica de tipos según especificaciones funcionales atacando en su lugar a una versión ingenua de la tesis de identidad psicofísica de tipos, basada en especificaciones fisicalistas, a la que nadie se esforzará en defender, conllevaría, como recuerda Bechtel (1988: 150), incurrir en la falacia del espantapájaros:

> Putnam […] imagines the brain-state theorist to claim that all organisms in pain –be they men, mollusks, Martians, machines, or what have you– are in some single common nondisjunctive physical-chemical brain state. Given the diversity of organisms, that claim *is* incredible. But the brain-state theorist who makes it is a straw man. (Lewis 1969: 233)

A juicio de Lewis, así pues, la teoría de identidad psicofísica nunca defendió una correspondencia biunívoca entre tipos de estados internos delimitados según criterios psicólogicos y tipos de estados internos delimitados según criterios físicos, aunque así pudieran darlo a entender los ejemplos que, con intención didáctica o retórica, acompañaban a sus primeras formulaciones. Al contrario,

> […a] reasonable brain-state theorist would anticipate that pain might well be one brain state in the case of men, and some other brain (or nonbrain) state in the case of mollusks. […] No mystery: that is just like saying that the winning number is 17 in the case of this week's lottery, 137 in the case of last week's. (Lewis 1969: 233)

Lo extraño, sin embargo, es que a Lewis no se le ocurriría afirmar que la propiedad denotada por el concepto "número premiado" sea una propiedad numérica, o, si se prefiere, matemática, aunque todos los entes que muestran esa propiedad sean de hecho números –o, más bien, sean analíticamente números, dada la intensión del concepto. Resulta obvio que la propiedad a la que apunta el concepto "número premiado" es de otro orden: no es una propiedad numérica o matemática la que engloba a todos y sólo los números premiados, como sí lo es en cambio la que engloba, por ejemplo, a todos y sólo los números menores de π. En realidad, debe no

serlo, salvo que nos forcemos a entender la propiedad de "guardar una relación aleatoria con el resto de términos de una serie numérica" como una propiedad matemática y asumamos la aun más inverosímil premisa de que tal relación determina unívocamente al sucesor para una serie dada, a fin de excluir de la extensión del concepto "número premiado" la infinitud de términos númericos que pese a satisfacer tal propiedad no han resultado premiados *de facto*. Cabe preguntarse entonces qué motivos podemos tener para admitir, como pretende Lewis, que la propiedad denotada por el concepto "dolor", o "ser un estado de dolor", sea una propiedad cerebral, cuando se nos concede de antemano no sólo que es falso que todos y sólo los estados de dolor tengan esa propiedad cerebral, sino que es falso también que tengan *alguna* propiedad cerebral. Si lo que engloba a todos los estados caracterizados por el concepto "ser un estado de dolor" es una propiedad funcional que puede darse en estados dispares desde el punto de vista neurofisiológico, e incluso en estados no caracterizables desde el punto de vista neurofisiológico, resulta un tanto obstinado insistir en que los estados de dolor, aquellos descritos por el concepto "ser un estado de dolor", son pese a todo estados cerebrales.

No mucho más convincente es el intento de Lewis de aliviar lo forzado de su posición. Es correcto, desde luego, en relación con el ejemplo del número premiado en un sorteo, que "[…] la aparente contradicción (una cosa idéntica a dos) se desvanece una vez que advertimos la relatividad al contexto implícita en uno de los términos del enunciado de identidad" (Lewis 1969: 233). Pero la conclusión a que eso nos aboca no es la que conviene a Lewis. A su juicio, de la misma manera que un mismo concepto de número premiado determina en distintos contextos distintas denotaciones, "[…] es el concepto fijo expresado en 'dolor' lo que determina el modo en que la denotación de 'dolor' varía según la naturaleza del organismo en cuestión" (Lewis 1969: 233); la moraleja en que debemos instruirnos es que:

> […t]he brain-state theorist cannot afford the old prejudice that a name of a necessary being (such as a state) must name it necessarily and independently of context. (Lewis 1969: 233)

Ahora bien: si el concepto fijo expresado en "dolor" es, como quiere Lewis, el concepto de un estado funcional, entonces no es cierto que dicho concepto "[…] determina el modo en que la denotación de 'dolor' varía según la naturaleza del organismo en cuestión". Lo que el concepto "dolor" determina es cuáles son las propiedades funcionales que un estado de un organismo o sistema debe mostrar para que posea la propiedad de ser un estado de dolor, cuál es el papel que dicho estado debe desempeñar en la organización funcional del organismo o sistema. En cambio, averiguar "[…] el modo en que la denotación de 'dolor' varía según la naturaleza del organismo en cuestión" es una tarea empírica diferente, que consiste en indagar qué propiedades físicas son características –si es que hay algunas que lo sean– de los estados que en distintos organismos se constituyen en estados de dolor por exhibir las propiedades funcionales requeridas para ello en virtud del concepto de dolor.

Dicho de otro modo: aunque contáramos con la completa certeza de que estados de dolor son aquellos que guardan tales o cuales relaciones con determinados estímulos, conductas y estados internos de un organismo o sistema, eso no nos diría nada sobre en qué estados físicos se encarna el dolor en organismos de una especie cuya fisiología ignorásemos. La razón es que el concepto de dolor, si efectivamente el dolor es un estado funcional, no determina cómo su denotación varía según la naturaleza de los organismos, sino más bien qué es lo que permanece inmutable a través de tales variaciones.

Probablemente, con todo, Lewis no quiere decir que el concepto de dolor determine *por sí mismo* cómo su denotación varía según la naturaleza del organismo, puesto que en tal caso la analogía le llevaría a sostener lo evidentemente insostenible: que el concepto de número premiado determina *por sí mismo* como varía su denotación en distintos sorteos, de suerte –nunca mejor dicho– que podríamos anticipar su denotación sin necesidad de esperar al desarrollo del sorteo. Obviamente, averiguar cuál ha sido el resultado de un sorteo es tarea empírica diferente de la de averiguar cuál es el concepto de número premiado.

Aunque "número premiado" no es en realidad un concepto funcional –no al menos en el sentido de denotar un estado interno funcionalmente caracterizado–, el paralelismo aún puede extenderse un poco más. La implementación física del dispositivo que genera el número, por ejemplo, es irrelevante en tanto en cuanto el carácter aleatorio de su registro de salida quede razonablemente garantizado: un bombo, una urna, una ruleta, cartones o barajas, una mente inocente o cualquier dispositivo electrónico de *aleatorización*, etc. Hay pues, aspectos relativos a la implementación física del dispositivo con el que se realiza el sorteo que el concepto de número premiado ni siquiera aborda. Pues bien, lo mismo vale para la indagación del concepto de dolor y de sus mudables encarnaciones físicas: la fórmula de determinación de la referencia del concepto de dolor –su caracterización funcional– excluye también aspectos relativos a la forma en que toma cuerpo el estado de dolor en diferentes casos, cuestiones que simplemente no se abordan en la caracterización funcional.

Por otro lado, la analogía con el caso del número premiado en un sorteo parece resquebrajar las conclusiones de Lewis más que fraguarlas. La propiedad expresada por el concepto "número premiado" no es una propiedad numérica, o matemática, precisamente porque los factores contextuales que determinan su referencia (los mismos que, *eo ipso*, nos permiten seleccionar extensionalmente todos y sólo los números premiados) no son de índole numérica o matemática. De la misma manera, parece que debemos concluir que la propiedad expresada por el concepto "dolor", o "ser un estado de dolor", no es una propiedad cerebral, puesto que los factores contextuales que determinan su referencia, y que nos permiten clasificar como tales a los estados de dolor y sólo a ellos, no son de índole neurofisiológica aunque su referencia sea, en algunos casos cuando menos, un estado cerebral: son, de acuerdo con el propio Lewis, de orden funcional.

La cuestión resulta algo menos oscura, en realidad, si se evita la analogía con el número premiado en un sorteo y su fatigoso desentrañamiento. Apenas unas pocas líneas antes de concluir que es el concepto fijo expresado mediante el término "dolor" lo que fija la referencia de dicho término en función de la naturaleza del organismo de que se trata, Lewis se había preguntado, retóricamente: "¿Por qué no concluir que el dolor = [el estado funcional] S_{17} = [el estado cerebral] B_{17} (en el caso de los humanos)?" (Lewis 1969: 233) La respuesta, a la vista de su conclusión, está al alcance de la mano. Pregunta por pregunta: ¿por qué no concluir además, si las premisas de que parte Lewis son correctas, que el dolor = el estado funcional S_{17} (en el caso general que determina ese "concepto fijo expresado mediante el término 'dolor'")? O bien: ¿por qué privarnos de ese concepto general de dolor si, de acuerdo con las tesis que Lewis y Putnam comparten, podríamos tenerlo? O, más apremiantemente: ¿qué justificación tendríamos, si no, para asegurar que empleamos el mismo concepto, y no meros homónimos, cuando usamos el concepto expresado mediante el término "dolor" a la hora de referirnos al dolor = S_{17} = B_{17} en el caso de los seres humanos tanto como a la hora de referirnos al dolor = S_{17} = B_{41} en el caso de –pongamos por caso– los cefalópodos? No podemos permitirnos el prejuicio –nos ha aleccionado Lewis (1969: 233, *supra*)– de pensar que el nombre de un estado mental debe nombrar ese estado mental, *ergo* funcional, de manera necesaria e independiente del contexto. De acuerdo, pero si disponemos, *ex hypothesi*, de nombres de estados mentales que hacen exactamente eso, ¿qué prejuicio nos fuerza a aferrarnos a otros que no lo hagan? O, dicho de otro modo, ¿qué prejuicio nos fuerza a aferrarnos a la ordenanza de que, porque ciertos nombres de estados cerebrales no nombren el mismo estado funcional de manera necesaria e independiente del contexto, tampoco puedan hacerlo los nombres de estados mentales que, contingentemente y en determinados contextos, compartan referencia con ellos?

Dar contestación a requerimientos de esta índole acabaría conduciendo a Lewis a tratar de articular los conceptos de estados psicológicos como conceptos disyuntivos, del estilo de S_{17}=(B_{17}∨B_{41}∨…), en una maniobra que Block y Fodor (1972a, *infra*) se esforzarían en recusar. No deja de resultar llamativo, en todo caso, el dictamen que el propio Lewis (1980) ofrecería años después respecto a la la viabilidad de un concepto disyuntivo de dolor tal que el dolor se identifique con (el estado funcional) S_{17} en el caso de los humanos normales y de organismos que padezcan dolor con diferentes fisiologías, *i.e.*, "dolor marciano", y con (el estado cerebral) B_{17} en el caso de los humanos normales y de aquellos humanos en los que el dolor venga ocasionado por estímulos y estados internos diferentes, u ocasione respuestas y estados internos diferentes, *i.e.*, "dolor insensato". A juicio de Lewis:

> This strategy seems desperate. One wonders why we should have a disjunctive […] concept of pain, if common men who suffer pain are always in pain according to both disjuncts […]. It detracts from the credibility of a theory that it posits a useless complexity in our concept of pain […]. (Lewis 1980: 217)

Las diferencias entre la estrategia disyuntiva descartada por Lewis y la que él mismo propugna son fáciles de alumbrar. Valgan como precipitada elucidación tres matices: uno –como el propio Lewis apunta– un mismo tipo de sujetos, los humanos normales, aparece en ambos términos de la disyunción en el primer caso, lo que no ocurre en el segundo; dos, a distintos (tipos de) sujetos corresponden caracterizaciones del dolor en diferentes vocabularios teóricos, en el primer caso, pero siempre en el vocabulario teórico de la neurofisiología, en el segundo; tres, la disyunción que Lewis considera parte de una estrategia desesperada no parece abocada a constar de más de dos términos, mientras que la que el propio Lewis impulsa tiene todos los visos de resultar, al menos en principio, una disyunción potencialmente infinita. Con todo, esas patentes diferencias no muestran por sí solas que una de las estrategias disyuntivas sea censurable y la otra plausible. Compete a Lewis convencernos, pues, de que identificar el dolor con la disyunción de los estados cerebrales en los que se encarna un determinado estado funcional no es, con sus propias palabras, una complicación inútil.

Sea como sea, Lewis se apresura a reparar –no podría ser de otro modo– en el flanco que Putnam (1967a: 229, *supra*) había dejado al descubierto en su defensa del funcionalismo ante el fisicalismo, al emplear contra el conductista lógico la distinción entre conceptos y propiedades. Basta eso, según estima Lewis, para establecer que los argumentos de Putnam (1967a) distan de haber demostrado que los estados mentales no puedan ser *tanto* estados funcionales *como* estados cerebrales, pese a que el concepto C_F de un estado funcional F y el concepto C_C de un estado cerebral C idéntico a F (F=C) no sean el mismo ($C_F \neq C_C$):

> The concept of any functional state as such does, of course, differ from the concept of any brain state as such. But Putnam is alive to the possibility that different concepts might be concepts of the same state; this observation is part of his own defense of the brain-state hypothesis against *a priori* objections. (Lewis 1969: 232)

Es en el contexto de esta escaramuza entre Putnam y Lewis donde surge la idea de una contraposición entre la tesis de identidad de estados funcionales, según la cual los estados mentales son eso, estados funcionales, y la tesis de identidad de especificaciones funcionales, según la cual los estados mentales son estados cerebrales especificados en términos funcionales. En palabras del propio Lewis (1969: 233):

> On this view, a functional state is better called a *functionally specified* state, and might happen to be a functionally specified brain state.

Se trata, claro, de la misma contraposición que Block (1980a: 35, 1997: 20; 1978: 67, *supra*) asimila a la que separa al funcionalismo *empírico* de Putnam, Fodor o Harman del funcionalismo *analítico* de Lewis, Armstrong o Smart. De acuerdo con Putnam o Fodor, el funcionalismo afirma la identidad entre tipos de estados mentales y tipos

de estados funcionales, incorporando la posible divergencia de implementación física de tales estados funcionales a la propia tesis de identidad, y entrañando con ello la falsedad del fisicalismo, entendido con alcance de tipos. La polémica entre la concepción del funcionalismo como *(i)* tesis de identidad de especificaciones funcionales y como *(ii)* tesis de identidad de estados funcionales puede verse al fin y al cabo como una cuestión de alcance lógico, cifrada en si hemos de concluir que cada tipo de estado mental se identifica con *(i)* un tipo de estados físicos que ocupa un determinado rol funcional, o más bien con *(ii)* un conjunto posiblemente dispar de estados físicos que tienen la propiedad funcional de ocupar el rol funcional característico de ese tipo de estado mental (*cf.* Schiffer 1986, *supra*).

El arbitraje de Block es tajante en este punto. A su entender, no sólo es falso a todas luces que, como Lewis solía reiterar, el funcionalismo avale un fisicalismo de tipos, sino que ni siquiera supone ningún aval para el fisicalismo de casos, dada –por decirlo a la manera de Block– su neutralidad ontológica. Por lo demás el fisicalismo de casos ni siquiera necesita –piensa Block– el presunto apoyo de argumentos de corte funcionalista. Sin embargo, el examen de la posición de Lewis esbozado por Block (1980a: 40, 1997: 20-21) contribuye a aislar el *locus* de la tensión, pero no la deshace. En efecto, para Lewis (1980), un análisis adecuado del concepto cotidiano de "sufrir dolor" lo interpretará como designador rígido de un estado funcional –el mismo en casos tan dispares como el nuestro y el de los miembros de una especie extraterrestre carente de sistema nervioso–; el análisis del concepto científico de "dolor", en cambio, revelará que se trata de un designador no-rígido del estado físico que desempeña un determinado rol funcional –distinto estado, claro está, en nosotros y en los extraterrestres en cuestión. Así que Lewis se ubica en el terreno del funcionalismo respecto de nuestros conceptos psicológicos cotidianos –es un funcionalista de corte analítico, o, como suele decirse, un funcionalista del sentido común– pero mantiene su lealtad al fisicalismo en el terreno de los conceptos teóricos desarrollados en la labor científica –rechaza, pues, el proyecto que Block denomina "psicofuncionalismo", en el que una laboriosa identificación entre estados mentales y estados funcionales viene a ser el fruto de la investigación empírica en psicología[176].

En cualquier caso, la neutralidad de las tesis funcionalistas respecto a la naturaleza de los estados mentales particulares no sólo arraiga en las reflexiones de Putnam sobre autómatas abstractos o de Fodor sobre la explicación psicológica, sino también, y hondo, en el análisis temáticamente neutral que Smart (1959) esgrimía en defensa de la teoría de la identidad psicofísica –"una caracterización funcional paradigmática" a juicio de Bechtel (1988: 150). Ya el propio Place (1956) se había esmerado en aclarar que la tesis de que los estados mentales –las experiencias de

[176] No es difícil entreoír en los alineamientos de Putnam, Fodor, Lewis y Armstrong en torno a esta cuestión los ecos de sus respectivas estirpes intelectuales: que Lewis y Armstrong, formados en el regazo del conductismo lógico y la teoría de la identidad psicofísica, se batan por la restitutición del fisicalismo con más denuedo que Putnam y Fodor, cuyo adiestramiento proviene sobre todo de la lógica y la lingüística, se prestaría sin duda a un detenido estudio desde la óptica de la sociología de la ciencia.

dolor, por reincidir en su perdurable ejemplo– son (idénticos a) estados o procesos cerebrales no conlleva la de que nuestra manera de describir dichos estados mentales sea idéntica a nuestra manera de describir los estados o procesos cerebrales con los que estos se identifican. Dado que, por regla general, admitir diferentes descripciones de una misma realidad no nos fuerza a admitir que nos hallemos ante diferentes realidades, la aclaración era a ojos de Place poco más que un ejercicio de explicitación del sentido común[177]. Poco después, Feigl (1958) invocaría la distinción fregeana entre el sentido y la referencia de un término para reforzar la posición de Place. La descripción cotidiana de una experiencia de dolor y su descripción física bien podían diferir en significado, pese a referirse a una única realidad, igual que "Héspero" y "Fósforo", o "el lucero del alba" y "la estrella vespertina", difieren en significado pese a aludir en todos los casos a un mismo cuerpo celeste, el mismo al que nos referimos como "Venus". Descubrir –tras una ardua labor experimental y teórica– que la experiencia de dolor es un determinado proceso cerebral bien puede constituir un avance científico de primera magnitud, y no una mera anécdota lexicográfica, igual que lo fue el descubrimiento de que era un mismo cuerpo celeste el que, en el firmamento boreal, era el primero en titilar sobre el horizonte nocturno y el último en desvanecerse con el alba[178]. Que esto sea un fenómeno de lo más corriente se deriva del hecho de que –como nos hiciera notar Frege (1892, *supra*)– el significado de un término no se agota en su referencia, sino que abarca también un determinado sentido, por lo que un enunciado de identidad entre términos correferenciales puede resulta informativo cuando los términos difieren en cuanto a su sentido. Así, aunque tanto "Héspero" como "Fósforo" se refieren a un mismo planeta, lo hacen bajo distintos sentidos: "Fósforo" lo aquilata según su trayectoria visible al amanecer, "Héspero" según su itinerario crepuscular. *Sinn* y *Bedeutung* servían en suma al fisicalismo moderno, en sus formulaciones pioneras, no sólo para amparar su hipótesis de que los conceptos que empleamos cotidianamente para describir nuestra vida mental designan de hecho procesos cerebrales, sino también, al

[177] Un ejercicio, por otra parte, que permite alumbrar los afluentes que el austero materialismo invocado por Place recibe de la remota tradición paralelista que nace en Malebranche. En un intento de dar respuesta a la posición epifenomenista ensayada por Huxley (1874) "[…] al hilo de la hipótesis de que los animales son autómatas", Conwy Lloyd Morgan (1896, 1900), por ejemplo, argumentaría que, siendo los fenómenos nerviosos y los de la vida consciente *concomitantes* y hallándose privados de comercio causal entre sí –cual era la tesis capital del paralelismo–, las descripciones subjetivas del psicólogo y las objetivas del fisiólogo son formas distintas de decir lo mismo. Hay, se diría, sólo un paso más en Place: puesto que así es, no hablamos ya de dos fenómenos concomitantes, sino de uno sólo.

[178] En un ensayo recogido en *Dimensions of Mind* –el mismo volumen que albergaba "Minds and Machines", de Hilary Putnam–, Feigl daría acaso la expresión más sumaria de su tesis: "[…]we may say that neurophysiological terms and the corresponding phenomenal terms, though widely differing in *sense* [...] do have identical *referents*" (Feigl 1960: 38). Conviene matizar, no obstante, a la luz de las palabras que siguen a ésas, el escueto materialismo que a menudo se atribuye a Feigl,: "I take these referents to be the immediately experienced qualities, or their configurations in the various phenomenal fields".

tiempo, para recalcar que la identidad de mente y cerebro es precisamente eso: una hipótesis. Tal como inicialmente lo había articulado Place (1956), siguiendo de cerca a Bertrand Russell (1903), la tesis de identidad psicofísica no se vale del "es" de la definición sino del "es" de la composición: no enuncia –diríamos– una leibniciana verdad de razón, sino una hipótesis empírica que, de ser verdad, será una verdad de hecho[179].

Reviste cierto interés evocar, al hilo de este asunto, la denuncia por parte de G.E. Moore de lo que él mismo tildó de "falacia naturalista" en el ámbito de la ética. En el contexto de esa denuncia, como quedó apuntado muy al comienzo de esta investigación, introduce Moore la cita de Butler –"Everything is what it is and not another thing"– que permearía buena parte del debate sobre el naturalismo no sólo en ética, sino también en epistemología y psicología. Al aparejar su aquilatada valoración del ataque de Moore, Prior (1949) nos recuerda que la falacia naturalista no sería sino:

> […] the assumption that because some quality or combination of qualities invariably and necessarily accompanies the quality of goodness, or is invariably and necessarily accompanied by it, or both, this quality or combination of qualities is *identical* with goodness. If, for example, it is believed that whatever is pleasant is and must be good, or that whatever is good is and must be pleasant, or both, it is committing the naturalistic fallacy to infer from this that goodness and pleasantness are one and the same quality. (Prior 1949: 1)

Es sumamente significativo que Prior opte por afinar la crítica del naturalismo forjada por Moore en términos de la distinción entre denotación y connotación, del mismo modo que Feigl (1958, *supra*) trataría no mucho después de pulir la tesis de identidad psicofísica recurriendo, aunque no de la mano de Mill sino de la de Frege, a la misma idea. En efecto, continúa Prior:

> What the man who commits the naturalistic fallacy fails to realize is that "good" and some other adjective may denote or be applicable to the same things, and yet not connote the same quality, *i.e.* describe the things in the same way. The difference between identity of denotation and identity of connotation may be brought out, as Professor Moore shows, by the following simple consideration: If the word "good" and, say, the word "pleasant" apply to the same things, but do not attribute the same quality to them, then to say that what is pleasant is good, or that what is good is pleasant, is to make a significant statement, however obvious its truth may appear to many people. But if the word "good" and the word "pleasant" not merely have the same application but the same connotation or "meaning" –if, that is to say, the quality of pleasantness is identical with the quality of goodness– then to say that what is good is pleasant, or that what is pleasant is good, is to utter an empty tautology, or, as Mill would call it, […] a "merely verbal" proposition; for

[179] La invocación a Frege sirve, pues, para recalcar el carácter empírico de la tesis de identidad psicofísica, pero –*nota bene*– no para sustentarlo, salvo que se adopte la problemática premisa auxiliar de que los términos que flanquean una definición comparten necesariamente referencia y sentido. Así parecen haberlo entendido, por ejemplo, Chacón y Rodríguez (2001), que exponen la distinción de Place entre definición y composición y el recurso de Feigl a Frege como cuestiones independientes.

> both statements are on this supposition merely ways of saying that what is pleasant is pleasant. (Prior 1949: 1-2)

De hecho, el propio Moore roza, sin desviarse un ápice de su trayectoria antinaturalista, la cuestión de la naturaleza de las sensaciones brutas, o *raw feels*, que sería capital en la transición del conductismo lógico al fisicalismo:

> Consider yellow, for example. We may try to define it, by describing its physical equivalent; we may state what kind of light-vibrations must stimulate the normal eye, in order that we may perceive it. But a moment's reflection is sufficient to show that those light-vibrations are not themselves what we mean by yellow. *They* are not what we perceive. Indeed, we should never have been able to discover their existence, unless we had first been struck by the patent difference of quality between the different colours. The most we can be entitled to say of those vibrations is that they are what corresponds in space to the yellow which we actually perceive. (Moore 1903: I, §10)

Algunos de los reparos que Prior opone al argumento de Moore encierran ya, veladamente, la cuidadosa reconstrucción de las tesis materialistas merced a la cual Place, Feigl y Smart le darían al fisicalismo un impulso tan poderoso, así como sus dificultades. La intuición de Moore es que la bondad es una propiedad última, elemental, que no es susceptible de quedar identificada con *ninguna otra*, simple o compleja. Pero esto –replica Prior (1949: 3) – no es decir mucho: "[…i]f we merely say that goodness is not identical with any other quality, this is itself a truism –it merely tells us that goodness is not identical with quality, simple or complex, with which it is not identical". Si en cambio, como Prior sensatamente supone, lo que Moore trata de mostrar es que la bondad no es idéntica a ninguna propiedad natural, entonces se verá forzado a encontrar otro modo de saldar la deuda de explicar qué entiende por propiedad natural que el de hacerlo por exclusión, sustrayendo del cómputo aquellas propiedades que, como bondad, toma por irreductiblemente éticas, puesto que en ese caso, por supuesto, habríamos regresado al terreno de la tautología. La identificación de dos propiedades –concluye provisonalmente Prior– no vulnera el por lo demás vaguísimo principio que Moore toma de Butler: así, por ejemplo, "[…] to say that goodness is pleasantness is not […] to deny that it is what it is, or to affirm that it is another thing –it is merely to deny that pleasantness is 'another thing'" (Prior 1949: 3). De la misma manera –se diría– que afirmar que un cierto estado mental es idéntico a un cierto estado cerebral no supone negar que sea el estado mental que es, ni afirmar que sea idéntico a algo distinto de sí mismo –sino únicamente negar que el estado cerebral en cuestión sea otra cosa, distinta del estado mental. Ahora bien, si la distinción entre propiedades éticas y propiedades naturales que Moore pretende reivindicar puede afincarse sobre terreno más firme que una estipulada exclusión mutua –como de hecho sucede, piensa Prior, tan pronto como tomamos en cuenta las diferencias intuitivas entre el dominio de los hechos y el de los valores–, entonces el argumento de Moore recobra, limadas sus aristas, buena parte de su fuerza. Lo que desenmascara, ciertamente, es el intento de ofrecer la identidad entre una propiedad

ética y una propiedad natural como una cuestión de análisis conceptual o una mera verdad de razón. Bien: ya Place (1956, *supra*) rechazaba contundentemente que la identidad psicofísica pudiera entenderse como identidad de definición. Esto, a juicio de Prior, es tanto como abandonar el naturalismo:

> […] an inconsistent ethical naturalist […] may clear himself of inconsistency […] by abandoning his naturalism –he may continue to insist that only pleasure, or conduciveness to survival, or whatever it may be, is good, but may preserve the significance of this assertion by sacrificing its certainty, admitting that its denial, though still in his opinion false, is not self-contradictory. (Prior 1949: 8-9)

Pero el dictamen resulta a todas luces demasiado taxativo: sólo puede armarse, de hecho, partiendo de definir la noción de naturalismo de modo que sólo cuenten como tales aquellas tesis que sostengan que su propia verdad es necesaria –que su contradicción es un sinsentido–, algo que, si bien podía resultar relevante en el contexto del naturalismo ético de la primera mitad del s. XX, dejaría de serlo en el del naturalismo psicofísico de la segunda. Como apunta el propio Prior:

> […] a naturalist can preserve his naturalism […] by admitting that the assertion that, say, pleasure and nothing but pleasure is good, is for him a mere truism; and that if Ethics be the attempt to determine what is in fact good, then the statement that what is pleasant is good is not, strictly speaking, an ethical statement, but only a way of indicating just what study is to go under the name of "Ethics" –the study of what is actually pleasant, without any pretence of maintaining that pleasure has any "goodness" beyond its pleasantness. (Prior 1949: 9)

Por el contrario, la escapatoria que Prior ofrece al naturalismo ético –escapatoria que, a su juicio, entrañaría el abandono del naturalismo– prefigura, como se ha dicho, el inminente derrotero del fisicalismo. Así visto, el giro decisivo en el caso del naturalismo psicofísico no fue otro que desligar la noción de identidad de las de definición y tautología, planteando la identidad entre estados mentales y estados cerebrales como una hipótesis empírico primero, o teórica, después. De esa manera, el dilema planteado por Prior al hilo de los argumentos de Moore se desvanece, y uno puede defender el fisicalismo sin obligarse a mantener que la propia tesis de identidad sea una especie de anexo analítico de la ciencia unificada de mente y cerebro, encargado de delimitar su objeto de estudio. Desde luego, habría sido raro que el intenso debate que alrededor de las tesis de Moore se libraba en Oxford, donde entonces estudiaban, no hubiera dejado su poso en el pensamiento de Place y Smart.

Al tiempo, la deriva eliminacionista que de manos de Feyerabend pronto emprendería el naturalismo psicofísico se ve redimida de antemano de la principal objeción con que Prior fustiga a su trasunto ético. El tono mesiánico que acabarían

por adquirir las proclamas de algunos eliminacionistas[180] quedó lúcidamente retratado *avant la lettre* por Prior:

> […] the naturalist] might add at the same time that he is not only not going to discuss goodness as a "non-natural" quality, but that in his belief there is no such quality, and that this is worth shouting from the housetops, as it liberates us from a transcendental notion which has haunted us too long. […] And such a man, it seems to me, should be prepared to state his position in an alternative way, namely, as a denial that there *is* such a study as Ethics –he should be prepared, for the sake of clarity, and to further the mental "liberation" in which he is primarily interested, to call his inquiry into the sources of pleasure, not Ethics, but some such name as "Hedonics"; or if he defines goodness as "conduciveness to survival", to call his substitute for Ethics "Biological Strategy". (Prior 1949: 9-10)

Dado que la identidad propuesta entre lo bueno y cierta propiedad natural se toma por una verdad de definición, dado que su negación no se pretende siquiera falsa sino carente de sentido, se torna difícilmente comprensible que expurgar nuestro pensamiento de tales nociones trascendentales resulte tan trabajoso. Como retóricamente se pregunta E.F. Carritt (1947: 33-34) –más conocido por sus trabajos sobre teoría estética– en la cita en que Prior (1949: 10) compendia esta rama de su argumentación: ¿podemos acaso ser liberados de una idea que ni siquiera podríamos haber albergado? Ahora bien, una vez que la identidad psicofísica ha quedado perfilada como una hipótesis empírica –o teórica– que se enfrenta en pie de igualdad conceptual con su rival antinaturalista, el horizonte de la eliminación se despeja. Cobra entonces sentido, al menos a primera vista, la labor de intentar liberarnos de una teoría falsa –como sugeriría Churchland (1988: *passim*; *cf.* por ejemplo 79) que es la psicología ordinaria– acerca de las causas de la conducta a la que, en nuestra ignorancia, nos aferramos.

Pero las fricciones entre el recurso a Mill por parte de Prior y el recurso a Frege por parte de Feigl no habrían de pasar inadvertidas. Cuando Feigl recurrió a la distinción entre sentido y referencia, hacía apenas unos años que Black y Geach habían dado a la imprenta un compendio, en inglés, de los ensayos de Frege, que aparecería en 1952. Era casi esperable que fuera Black, entonces, quien planteara a Smart la objeción de que la teoría de la identidad psicofísica estaba forzando el espíritu y la letra de la semántica fregeana. Al distinguir el sentido y la referencia de un término, Frege había enfatizado que lo que recogemos bajo el rubro de "sentido" no es sino el "modo de presentación" –o, más bien, de determinación: *cf. supra*– de la referencia. Pero para que la noción de *Sinn* llegue a consumar la tarea que Frege le encomienda es preciso convenir en que se trata de una propiedad capaz de ser, simultáneamente, aprehendida por la mente e instanciada por un objeto particular. La propiedad de ser el último cuerpo celeste en desprenderse de las madrugadas septentrionales, por ejemplo, es una propiedad que Venus instancia, y que nuestra

[180] *Cf.*, por ejemplo Churchland (1984/1988), especialmente los párrafos acerca de "La expansión de la conciencia introspectiva" del último capítulo, como Churchland (1984/1988: 256, *supra*).

mente de algún modo apresa cuando nos referimos al lucero del alba; es, obviamente, una propiedad distinta de la que configura nuestros pensamientos sobre la estrella vespertina. Si regresamos, con estas consideraciones presentes, a la cuestión de la identidad entre la mente y el cerebro, toparemos con la conclusión hacia la que Black nos apremiaba: decir –por ejemplo– que el dolor es un determinado proceso cerebral, pero que el giro cotidiano "sentir dolor" y el enunciado fisicalista que designa dicho proceso difieren en cuanto a su sentido, es *eo ipso* reconocer que el dolor y el proceso cerebral difieren en al menos una propiedad, aquella que constituye su modo de presentación. Ahora bien, no cabe al dolor en nuestro ajetreo diario –ajeno al de la investigación neurológica– otro modo de presentación que el de dolor sentido, sufrido o sobrellevado; en el caso de los términos con los que solemos hablamos del dolor, como de los estados psicológicos en general, el sentido se instancia –así pues– en propiedades irreduciblemente mentales. Por mucho que el dolor se identifique con un proceso cerebral, habrá que adjudicar a ese proceso –piensa Black– una naturaleza mental pareja a su naturaleza física.

Pues bien, el análisis temáticamente neutral del vocabulario mentalista tenía por propósito desmantelar las críticas de Black. El truco consistía en articular un modo de referirnos a nuestros estados mentales que –a diferencia, como había advertido Black, de los giros mentalistas del lenguaje ordinario– no supusiera compromiso alguno respecto a que su naturaleza fuese mental o física (o sea, que fuera temáticamente neutral), pero que, pese a ello, preservara íntegro el significado de tales giros. En 1959 Smart se mostraba convencido de que semejante proeza conceptual era factible: bastaría con reinterpretar –con el ejemplo que retoma Rabossi (1995: 25)– "Veo una postimagen anaranjada" como "*Algo acaece que es como lo que acaece* cuando tengo mis ojos abiertos y hay una naranja bien iluminada frente a mí". La tesis de identidad psicofísica, entonces, podría quedar reformulada haciendo equivaler el referente de esas expresiones reformuladas al de las descripciones neurológicas que fueran proporcionando las ciencias del cerebro, y la diferencia de sentido entre unas y otras no remitiría a propiedades inherentemente mentales que arruinaran el temple fisicalista del proyecto. Y como las desgarbadas expresiones temáticamente neutrales recogían plenamente el significado de los conceptos psicológicos del sentido común, la pretensión de estar proporcionando un análisis hecho y derecho de dichos conceptos quedaría reivindicada. Se trataba, en síntesis, de mostrar que, a la hora de fijar el sentido en el que nos referimos cotidianamente a nuestros estados psicológicos, la apelación a propiedades mentales es un trámite prescindible.

La táctica de Smart, una vez más, había quedado prefigurada en el trabajo de Place (1956). Al impugnar la falacia fenomenológica –la creencia de que al describir nuestros estados mentales atribuimos propiedades a sustancias u objetos mentales–, Place proponía esquivarla analizando, por ejemplo, "Siento un sabor dulce" como "Tengo una experiencia gustativa que es del mismo tipo que las que suelo tener cuando paladeo algo dulce", en lugar de como "Tengo una experiencia gustativa que

es dulce". Esta última lectura, que atribuye a la experiencia de dulzor la propiedad de la dulzura, invalidaría, en efecto, la tesis de identidad, pero resulta –si la denuncia de Place es robusta– tan falaz como eludible.

La acuciosa búsqueda de la neutralidad temática no tardó en cosechar acerbas críticas. Un argumento de Cornman (1962), que Rabossi (1995) ensaya sucintamente, era particularmente vigoroso: si las expresiones de experiencia fenoménica y sus transcripciones temáticamente neutrales eran en efecto sinónimos, responderían a la reciprocidad y transitividad que caracteriza a la relación de sinonimia, pero incluso si aceptamos que decir (*i*) "Veo una postimagen anaranjada" vale tanto como decir (*ii*) "*Algo acaece que es como lo que acaece* cuando tengo mis ojos abiertos y hay una naranja bien iluminada frente a mí", no podremos admitir que decir (*iii*) "Veo una forma redonda", que en el mundo fenoménico del sujeto queda perfectamente descrita por (*ii*) "*Algo acaece que es como lo que acaece* cuando tengo mis ojos abiertos y hay una naranja bien iluminada frente a mí", sea sinónimo de (*i*) "Veo una postimagen anaranjada" –o sea, que *(i) = (ii)* y *(ii) = (iii)*, pero *(i) ≠ (iii)*. Por consiguiente, ni las alambicadas locuciones pergeñadas por Place y Smart son sinónimas de las expresiones coloquiales de experiencias conscientes, ni apresan la anhelada neutralidad temática, y –lo que es más grave– "[…] cualquier maniobra tendente a eliminar la dificultad incluirá un rasgo tópico [temático] asociado a la experiencia del agente." (Rabossi 1995: 25)

A juicio de algunos críticos del fisicalismo, de hecho, el argumento de Black conserva casi intacto su vigor a día de hoy; es el caso, por ejemplo, de White (1986, 1999). Sea como sea, lo que interesa en este punto es recobrar la idea de que la neutralidad ontológica del funcionalismo estaba ya presagiada en los trabajos de los teóricos de la identidad psicofísica. Desde luego, si hay verdad en la tesis de que las expresiones temáticamente neutrales de Smart (1959) no entrañan mención alguna a propiedades mentales en la determinación de su sentido, la habrá, *a fortiori*, en la tesis pareja de que tampoco implican mención alguna de propiedades físicas. Este es, de hecho, un resultado apetecido por Place, ya que de lo contrario su concepción de la identidad psicofísica como una verdad contingente quedaría gravemente dañada, si no del todo exánime. Comoquiera que Smart, en cambio, consideraba que la tesis de identidad psicofísica era una hipótesis de orden teórico más que empírico –cuyo respaldo no habría de encontrarse en la observación, que nunca podría desechar el epifenomenismo, sino en el principio de parsimonia (Smart 1959)–, quizá hubiera podido permitirse que la presunta neutralidad temática de su análisis acabara ladeándose hacia el fisicalismo. Pero no fue así: "estados mentales del mismo tipo que los que suelo tener cuando…" tanto podrían darse –hasta donde el análisis de la expresión autoriza a establecer– en organismos dotados de intrincados sistemas nerviosos, como en seres con una anatomía radicalmente distinta, en criaturas angelicales o en inexorables máquinas.

La controversia, en este punto, vuelve a transitar por un terreno ya explorado en relación con los vínculos entre fisicalismo y conductismo lógico: el de la

refractariedad del funcionalismo a los planteamientos reduccionistas. Según veíamos, el hecho de que el funcionalismo aspire a caracterizar cada tipo de estado o proceso mental en un vocabulario que, aparte de términos lógico-matemáticos, incluya únicamente descripciones de aferencias, eferencias y otros estados internos lo torna engañosamente afín al conductismo lógico; engañosamente, porque la clave es que el funcionalismo logra su objetivo –si lo logra– merced a la cuantificación sobre estados y procesos físicos *en tanto que* encarnación de estados y procesos mentales. Dicho de otro modo: en la caracterización de un tipo de estado mental aparecen términos que designan estados del sistema y que quedan caracterizados funcionalmente –como quien dice– en otro fragmento de la teoría. Así, el peso de lo mental es sustentado solidariamente por la totalidad de la teoría, sin llegar nunca a recaer sobre ninguno de sus fragmentos –sobre ninguna de las caracterizaciones de tipos de estados mentales en particular. Pero ahí permanece: que todas las partes de una estructura contribuyan a soportar una carga no implica que no haya tal carga. Recuérdese, al hilo de la valoración de los compromisos ontológicos que conlleva el formalismo de Ramsey, la conclusión de Block:

> It is a simple fallacy to suppose that if each mental term is defined in terms of the others (plus inputs and outputs), then each mental state is defined nonmentalistically. (Block 1978: 78, *supra*)

Someramente intenta Block aplicar el mismo razonamiento contra la constelación teórica formada por funcionalismo analítico, análisis temáticamente neutral y fisicalismo de tipos. A su entender, aun si –de la mano de Lewis– diéramos por buena la tesis de que del análisis del significado de los términos psicológicos del discurso ordinario se sigue que un estado mental *es* un estado que ocupa un determinado papel causal, no estaríamos por ello forzados a admitir que un estado mental sea el ocupante de un determinado papel causal *especificado en términos no mentales* –sino temáticamente neutrales, se entiende. Es difícil argumentar lo contrario, pero Block parece pasar por alto que podríamos tener motivos independientes para aceptar el análisis funcionalista de nuestra jerga psicológica *y* para aceptar su análisis temáticamente neutral; sin duda tendríamos, en tal caso, motivos lógicos para conjugar ambos análisis y posarnos suavemente sobre la conclusión que Block trata de esquivar. Si, por ejemplo, "dolor", tal como usamos cotidianamente el término, resultara referirse al estado que ejerce cierto papel funcional, *y* "dolor", tal como usamos cotidiamente el término, resultara adecuadamente traducible a un vocabulario temáticamente neutral, en la línea ingeniada por Smart, entonces seguramente sí tendríamos buenas razones para convencernos de que el dolor es el ocupante de cierto papel funcional especificado de forma temáticamente neutral –aunque fuera sólo por transitividad. A lo que Block necesita poner peros, si ha de perseverar en su concepción antirreduccionista del funcionalismo, es por tanto a la eficacia del análisis tematicamente neutral, pues obstaculizar la del análisis funcionalista sería tanto como renunciar a dicha

concepción. Ahora bien, parece sensato aventurar que el razonamiento esgrimido contra la asimilación de funcionalismo y conductismo lógico menoscaba también la credibilidad del análisis temáticamente neutral como herramienta reduccionista. De la misma manera que la lectura funcionalista de los términos psicológicos parece, engañosamente, garantizar la posibilidad de prescindir de ellos –y sin embargo, al cuantificar sobre ellos, los hace irrenunciables–, el análisis temáticamente neutral parece ofrecernos un modo de caracterizar los estados mentales –de cara, en el trabajo de Smart, a su posterior identificación con estados físicos– sin mencionar sus propiedades mentales, pero la referencia a estados mentales queda de hecho irremisiblemente entrañada en el análisis. So pena de seguir abusando del ejemplo: que el dolor se identifique como aquel estado que es efecto de los estímulos $E_1 \ldots E_n$, y causa de las conductas $C_1 \ldots C_n$ así como del estado S_i, no puede contar como análisis reduccionista en la medida en que S_i (que equivale –pongamos por caso– a "preocupación") se identifique a su vez mediante el mismo expediente. La identificación, entonces, del dolor como un estado indistinguible de aquel en el que me encuentro al confrontar los estímulos $E_1 \ldots E_n$, etc., a la moda temáticamente neutral, bien contendrá una mención irrevocable de S_i , bien, si no la contiene, será sencillamente falsa, suponiendo –claro– que el análisis funcional sea correcto. Así que Block, aunque aparente errar levemente la trayectoria de su ataque, lleva las de ganar en esta contienda. Parece pues de rigor, después de todo, que en la misma vena en la que ha rechazado la interpretación del funcionalismo como una variante del conductismo lógico (Block 1996: 17, *supra*), se opusiera también a la tesis de que el análisis temáticamente neutral convierte al funcionalismo en una variante del fisicalismo de tipos.

La posición de Fodor (1974), por supuesto, es que hay motivos sólidos para entender el funcionalismo como una tesis de identidad de estados funcionales irreconciliable con el fisicalismo de tipos, motivos que se cifran ante todo en las notables ventajas de índole epistemológica que esto aportaría:

> The reason it is unlikely that every kind corresponds to a physical kind is just that (a) interesting generalizations… can often be made about events whose physical descriptions have nothing in common; (b) it is often the case that whether the physical descriptions of the events subsumed by such generalizations have anything in common is, in an obvious sense, entirely irrelevant to the truth of the generalizations, or to their interestingness, or to their degree of confirmation, or, indeed, to any of their epistemologically important properties; and (c) the special sciences are very much in the business of formulating generalizations of this kind. (Fodor 1974: 103)

Poco antes, en su trabajo conjunto de 1972, Block y Fodor habían dado por buenas las "consideraciones empíricas" (Block y Fodor 1972a: 46) de Putnam (1967a) que habrían descuadernado a su entender el fisicalismo de tipos. Entre ellas, en primer lugar, la concepción de la función cerebral de Lashley (1929), que impregna así

nuevamente los orígenes del funcionalismo y el cognitivismo[181] –y con ello, de la mano de Jean Pierre Flourens (1824), un cartesiano hecho y derecho, tan conocido por sus diatribas contra la frenología (Flourens 1842) como contra el evolucionismo (Flourens 1864), los enlaza una vez más con la tradición racionalista. En efecto, la doctrina de equipotencialidad propugnada por Lashley y los indicios de readaptación de la topografía funcional del cortex en ciertos casos de lesión cerebral sobre los que dicha doctrina se apoyaba, y cuya formulación primera es mérito de Flourens, aparecen como un signo claro de que una interpretación lógicamente fuerte de la tesis de identidad psicofísica sería –digámoslo así– empíricamente débil[182]. Pero junto a Lashley, esta vez, se alista también Darwin y la convicción de que las semejanzas morfológicas aparentes entre distintas especies podían ocultar fisiologías profundamente diferenciadas, pero que respondieran a presiones evolutivas similares; trasladar tal noción de convergencia al caso de semejanzas *conductuales* aparentes era casi rozar la idea de realizabilidad múltiple. Por último, la mera "[…] posibilidad conceptual de que se pudiera aplicar predicados psicológicos a artefactos" (Block y Fodor 1972a: 46) se perfilaba como razón suficiente para abandonar el fisicalismo de tipos. No habían irrumpido todavía en el debate los alienígenas que con tanta naturalidad lo habitarían a partir de Lewis (1980), y que Horst (1996) –en el contexto de una afilada crítica de la teoría computacional de la mente– añade a su recapitulación de las ganancias que el funcionalismo ofrecía en comparación con un fisicalismo de tipos, con el cual también él lo considera incompatible:

> Additional arguments for the benefits of functionalism over reductionism were marshaled on the basis of Lashley's thesis of equipotentiality, the convergence of morphological and behavioral features across phylogenetic boundaries, and the possibility of applying psychological predicates to aliens and artifacts. (Horst 1996: 51)

Intentar volver a Lewis (1980) y al dolor marciano contra sí mismo es, en todo caso, una astucia que cuenta con el refrendo –irrenunciablemente burlón, eso sí– del propio Fodor (1985). El funcionalismo respaldaba a su juicio –recordemos:

> […] the emerging intuition that the natural domain for psychological theory might be physically heterogeneous, including a motley of people, animals, Martians (always, in the philosophical literature, assumed to be silicon based), and computing machines. (Fodor 1985: 15, *supra*)

Después de auditar la contabilidad explicativa de ambas concepciones del funcionalismo, Levin (2004), por su parte, presenta un balance equilibrado, en el que los débitos y haberes de cada una se arquean –como quien dice– contrapuestos. La

[181] De hecho, Fodor ya citaba a Lashley en las últimas páginas de *La explicación psicológica*, también con propósito de hallar respaldo contra el reduccionismo psicofísico en la tesis de equipotencialidad (Fodor 1968: 185).

[182] Pero *cf.* Shapiro (2000, 2004) o Wilson y Craver (2007) sobre von Melchner *et al.* (2000), *infra*.

estrategia fraguada por Putnam y Fodor era identificar la propiedad de atravesar un determinado (tipo de) estado psicológico con una propiedad funcional, de segundo orden, que puede quedar instanciada en multitud de (tipos de) estados en virtud de sus propiedades de primer orden –hasta donde sabemos, pero no necesariamente, físicas. Ello implica obviamente enfatizar que es la presencia de la propiedad funcional en cuestión –o si se prefiere, el tipo de estado funcional que determina–, pero no el tipo de estado físico en el que quede instanciada en un conjunto de casos o en otro, la que constituye un tipo de estado psicológico. Gracias a ello, la tesis de identidad de estados funcionales permite, como venimos viendo, dar cuenta con mayor elegancia de la posibilidad de que se den estados psicológicos del mismo tipo en sistemas cuya estructura física sea indefinidamente diferente –o incluso, si se quisiera, en sistemas cuya estructura no sea física. Por el contrario, Lewis y Armstrong se inclinan por identificar los (tipos de) estados psicológicos con los (tipos de) estados cuyas propiedades de primer orden –hasta donde sabemos, físicas– determinen que venga instanciada, en un conjunto de casos o en otro, la propiedad funcional, de segundo orden, relevante. O mejor, deshaciendo la aparente asimetría respecto a si la tesis funcionalista se predica de estados psicológicos o de la propiedad de hallarse en dichos estados: la propuesta de Lewis y Armstrong pasa por identificar la propiedad de atravesar un determinado (tipo de) estado psicológico con un conjunto de propiedades físicas de primer orden (o con la propiedad física que éstas forman en su conjunto) instanciadas en un determinado (tipo de) estado físico que, en un conjunto de casos o en otro, instancia también, en virtud precisamente de instanciar esas propiedades físicas de primer orden, una determinada propiedad funcional de segundo orden. En consecuencia, lo que se enfatiza es que un tipo de estado psicológico no es sino la presencia de una propiedad física –o el tipo de estado físico que determina esa propiedad. De ese modo –es cierto– la tesis de identidad de especificaciones funcionales se ve abocada a explicaciones más prolijas de posibles coincidencias de tipos de estados psicológicos en distintos tipos de sistemas físicos, orgánicos o inorgánicos –y, de paso, excluye la posibilidad de estados psicológicos en sistemas que no sean físicos–, pero a cambio adquiere primacía en un terreno crucial: el de la causalidad. En efecto, si un estado psicológico *es* la presencia de una propiedad física, entonces resulta meridianamente claro, a primera vista, cómo un estado psicológico puede causar una respuesta muscular que se traduzca en una conducta, cómo puede venir causado en último término por el efecto de patrones ambientales de energía física sobre las células de los órganos sensoriales, o cómo puede ser la causa o el efecto de otros estados psicológicos:

> Taking mental states to be the first order state-types which, in each species, satisfy the functional definitions makes it possible to say, literally, that *pain* (given the presence, or absence, of certain other mental states) *causes wincing*, whereas on the view that pain is a second-order property partially defined by its tendency to produce wincing, one can say only that wincing is a *manifestation* of pain. (Levin 2004: §3.4)

Acaso sea posible, también, reconstruir la noción de causa y aplicarla a los estados psicológicos bajo el presupuesto de que un estado psicológico es la presencia de una propiedad funcional, de segundo orden, pero tal reconstrucción arrojará sin duda una noción de causalidad menos llana e intuitiva. Éste es el exacto contrapeso de las dificultades de la tesis de identidad de especificaciones funcionales para vérselas con la naturaleza de los estados psicológicos *per se*, en tanto que estados que seres físicamente diferentes pueden compartir. En síntesis:

> Functional specification theories and F[unctional] S[tate] I[dentity] T[heorie]s appear, that is, to have different strengths and weaknesses: straightforwardness of causal explanation versus universality of application. (Levin 2004: §3.4)

No es de extrañar, pues, que desplegar una interpretación convincente de la noción de causalidad que resulte aplicable a estados psicológicos haya sido uno de los quehaceres en los que Fodor se haya volcado con mayor devoción, máxime teniendo en cuenta que él mismo consignaba "la autonomía ontológica de los particulares mentales y, con ello, el carácter causal de la interacción entre la mente y el cuerpo" (Fodor 1981a: 9, *supra*) como el legado primordial que el funcionalismo habría logrado salvar del naufragio de la teoría de identidad psicofísica.

Pero la interpretación del funcionalismo como una tesis de identidad de especificaciones funcionales, propugnada por Smart, Armstrong y Lewis, esconde una tara más dañina que la dificultad para generalizar entre especies o tipos de sistemas, o –mejor dicho– una variante sensiblemente más virulenta de esa misma dificultad. El fisicalismo de tipos, también en sus formulaciones de inspiración funcionalista, subestima la potencia de la intuición según la cual el mismo tipo de estado mental puede venir instanciado por estados con propiedades neurológicas (y, en último término, físicas) dispares cuando restringe esta posibilidad al caso de especímenes de distintas especies, o de máquinas, que –supongamos– compartan nuestros estados psicológicos. La intuición funcionalista ataca también, aunque sea quizá con ímpetu algo mitigado, al presupuesto de que el mismo tipo de estado mental haya de quedar instanciado por el mismo tipo de estado neurológico (o físico) en distintos individuos de la misma especie, o incluso en el mismo individuo en distintos momentos. A esto último apunta certeramente Rabossi (1995) al anotar que:

> […e]l cerebro tiene una plasticidad tal que la identificación de tipos [de estados] psicológicos con tipos neurofisiológicos resulta prácticamente imposible. Respecto de una misma persona, cabe dudar de que el evento neurofisiológico en que consiste su sentir dolor en un momento dado […] sea el mismo en el que consiste su dolor en otro momento. (Rabossi 1995: 31)

La razón, sin embargo, no estriba sólo en la mayor o menor labilidad del sistema nervioso, sino, cardinalmente, en la extrema finura y poder de abstracción del aparato de descripción de estados mentales que es uno de los dones del lenguaje.

Resulta poco creíble, verbigracia, que el pensamiento de que mañana hará buen tiempo sea identificable, en todas y cada una de sus manifestaciones a lo largo de la vida de cualquiera de nosotros, por unas propiedades neurológicas (o físicas) que lo distingan de cualquier otro pensamiento, y que además habrán de ser las mismas en todos nuestros congéneres[183]. Incluso forzando un poco menos la resolución de nuestros mecanismos de descripción de estados psicológicos, resulta poco creíble que las creencias, en todas y cada una de sus manifestaciones a lo largo de la vida de cualquiera de nosotros, sean discernibles de los deseos, y estos de los anhelos, en virtud de propiedades neurológicas (o físicas) que les sean privativas. Pero si esto es así, entonces el conjunto de sujetos al que la tesis de identidad de especificaciones funcionales debe acotarse no será ni siquiera la especie: quizá sea el individuo para algunos estados o procesos psicológicos, o alguna segmentación temporal suya para otros, acaso –quién sabe– haya a menudo que ligar la explicación a grupos de individuos delimitados según criterios culturales, sociales, etc. Con ello, la tesis de identidad de especificaciones funcionales habrá dejado de parecerse a una generalización científica, no sólo por su tosca prolijidad sino, sobre todo, porque habrá quedado desprovista casi por completo de su poder explicativo y predictivo, o, al menos, de su radio nomotético –igual que veremos, *infra*, al hilo del escrutinio de los argumentos de Kim contra la lectura antirreduccionista del funcionalismo.

Los argumentos de Block y Fodor (1972a) contra el funcionalismo de tabla de máquina iban precedidos de la convicción de que tanto el conductismo lógico como el fisicalismo de tipos –cuya inviabilidad se despachaba en unas pocas líneas– estaban epistemológicamente desahuciados. Aun así, es llamativo que en el expediente de desacreditación del fisicalismo de tipos, el argumento decisivo se fundara –siguiendo a Putnam (1967a)– sobre:

> […] the empirical likelihood that creatures of different composition and structure, which are in no interesting sense in identical physical states, can nevertheless be in identical psychological states; hence that types of psychological states are not in correspondence with of physical states. (Block y Fodor 1972a: 46)

Desde luego, establecer que la identidad psicofísica de tipos sea una mala apuesta empírica también para distintos individuos de la misma especie, o el mismo individuo en distintos momentos, no es tan sencillo como hacerlo para el caso de "criaturas con diferente composición y estructura" –dicho de otro modo, que se diera identidad psicofísica de tipos intraespecífica no nos dejaría tan estupefactos como descubrir la variedad de identidad psicofísica que Block y Fodor descartan. Cierta merma en el empuje de la objeción viene compensada –sin embargo– por una trayectoria prolongada: nos es dado desvencijar ahora cabalmente no sólo la tesis de identidad psicofísica de tipos, sino también la tesis de identidad de especificaciones

[183] La misma consideración se contrapondrá, *infra*, al énfasis de Bechtel y Mundale (1999) en la pobre resolución de la descripción de capacidades psicológicas, que ellos contrastan con la minuciosidad de la descripción neurofisiológica para tratar de desacreditar la tesis de realizabilidad múltiple.

funcionales. De ahí que resulte raro que Block y Fodor pasaran por alto el asunto, sobre todo teniendo en cuenta que su trabajo comenzaba con la constatación de que:

> As far as anyone knows, different organisms are often in psychological states of exactly the same type at one time or another, and a given organism is often in psychological states of exactly the same type at different times. Whenever either is the case, we shall say of the psychological states of the organism(s) in question that they are *type identical*. (Block y Fodor 1972a: 45)

Todavía en su influyente trabajo contra la doctrina positivista de la unidad de la ciencia, Fodor (1974) sólo diferencia implícitamente entre una versión de la tesis de realizabilidad variable que concierna a organismos o sistemas de diferente tipo –tipo biológico o, en algún sentido, físico, se entiende– y otra que ataña a un único organismo o sistema en distintos momentos.

Las repercusiones de la distinción han sido estudiadas *inter alia* por Bickle (1998, 2006), que mantiene un severo escepticismo respecto a la propia tesis de realizabilidad múltiple. A su juicio, la radicalización del sentido de la tesis de realizabilidad múltiple que tiene lugar al aplicarla a los distintos estados mentales del mismo tipo que pueda atravesar un organismo a lo largo del tiempo ha convertido el argumento de Putnam (1967a) no en una pretendida refutación de la identidad psicofísica, sino en una reivindicación del carácter irreductible de la explicación psicológica con respecto de las ciencias naturales, o físicas, que –estima– es hegemónica en la reciente filosofía de la mente angloamericana (Bickle 2006: §1.5). Salvo a modo de hipótesis histórica, sin embargo, es difícil ver qué sustento pudiera darse a la idea de que la realizabilidad múltiple en un único organismo pueda respaldar argumentos antirreduccionistas de alcance general, mientras que la realizabilidad múltiple en distintos tipos de organismo sólo alcance a un argumento antifisicalista. Parece, más bien, que la realizabilidad múltiple, de verificarse para diferentes tipos de organismos[184], obliga al fisicalista –o, cuando menos, al defensor de la tesis de identidad psicofísica– a refugiarse, como Lewis (1969), en reconstrucciones funcionalistas de su posición, en las que el vocabulario neurofisiológico aparece sólo en enunciados disyuntivos; si la realizabilidad múltiple se viese empíricamente contrastada para los mismos organismos en diferentes momentos, el cobijo de la reconstrucción disyuntiva quedaría devastado –a menos que accediéramos a incluir tantos términos en la disyunción como ocasiones en las que diversos organismos o sistemas albergaran o pudieran albergar el estado mental en cuestión. Como certeramente anota Bickle (2006: §1.1), la tesis de realizabilidad múltiple para idéntico sistema a través del tiempo es más radical "[…] because there could be a disjunction of physical states realizing each mental kind for every existing cognizer".

[184] Una reflexión popperiana: la tesis de realizabilidad múltiple puede verificarse, y no sólo ser sometida a intentos de falsación, en la medida en que no es una generalización universal, sino la negación de una.

El caso es que, al igual que hay indicios empíricos de la verosimilitud de la tesis de realizabilidad variable para distintos tipos de organismo o sistema –Bickle (2006: §1.2), por ejemplo, menciona "[…] comparative neuroanatomy and physiology, facts about convergent evolution, and the corticalization of function (specially sensory function) as cortical mass increases across species"–; los hay también que –ya sabemos: desde Flourens, desde Lashley– hablan de la veracidad de su interpretación para idéntico organismo o sistema en distintos momentos. Entre las constataciones más tempranas de tales indicios destaca la de Endicott (1993), que conjura en su argumentación la capacidad de "[…] distinct neural structures and processes to subserve a given psychological function owing to trauma, damage, changing task demands, development, and other factors" (Bickle 2006: §1.5; *cf.* Block y Fodor 1972a: 46 y Horst 1996: 51, *supra*)). Entre las formulaciones más contundentes de las consecuencias de estos hechos –si hemos de conceder su interpretación antifisicalista–, el propio Bickle (2006: §1.5) destaca la de Horgan (1993: 308):

> Multiple realizability might well begin at home. […T]he intentional states we attribute to one another might turn out to be radically multiply realizable at the neurobiological level of description, *even in humans*; indeed, even in *individual humans*; indeed, even in an individual human *given the structure of his central nervous system at a single moment of his life*.

De hecho, el razonamiento esgrimido por Block y Fodor para acallar –según creen– los últimos aleteos del reduccionismo transitará, aunque sin llegar a plantearla abiertamente, aún más cerca de la idea de que las propiedades neurológicas de distintas instancias del mismo tipo de estado mental puedan ser diferentes incluso dentro de los márgenes de la especie, o del individuo. El adversario al que se ha de plantar cara es una reconstrucción disyuntiva de las tesis fisicalistas (o conductistas) de acuerdo con la cual cada tipo de estado psicológico queda identificado con una disyunción de tipos de estados físicos (o de conductas o disposiciones a la conducta, si es el conductismo lo que tratamos de rescatar), de tal modo que albergar un estado psicológico equivale a quedar adecuadamente descrito por uno u otro de los términos de la disyunción[185].

Pero ya Block y Fodor (1972a) advirtieron de las dificultades que erizan esa ruta. En primer lugar, sería un error dar por sentado que exista siquiera una disyunción de estados físicos, o de disposiciones, que sea distintiva de cada estado psicológico:

> For example, there is really no reason to believe that the class of types of behaviors which, in the whole history of the universe, have (or will have) expressed rage for some organism or other, is distinct from the class of types of behaviors which have expressed, say, pain. (Block y Fodor 1972a: 47)

[185] La estrategia emprendida por Lewis (1969), pese a su impugación por parte de Block y Fodor, ha sido prolífica, y alienta de un modo u otro las propuestas de Kim (1972, 1989, 1992), Lycan (1981), Rabossi (1995), Chalmers (1996), Jackson (1998), Dennett (2001) o Noë (2004).

Lo único que podríamos hacer corresponder, entonces, a esa disyunción de tipos de conductas (o, *mutatis mutandis*, de estados neurológicos) es una disyunción de estados psicológicos. Naturalmente, identificar disyunciones de tipos de estados psicológicos con disyunciones de tipos de estados físicos no representa un gran avance explicativo, máxime cuando, en principio, no parece haber restricciones respecto a qué términos de una disyunción puedan aparecer también en otras disyunciones del mismo tipo, que se identifiquen con otras disyunciones del tipo opuesto. Aunque Block y Fodor no hurgan en la herida, porfiar en esta vía puede a la larga encaminarnos a forzar en exceso la noción de identidad, aproximándola a conclusiones lisamente contradictorias: la tesis acaba siendo parecida a la de quien asegurara que cada tipo de estado psicológico es idéntico a *algún* tipo de estado físico que, sin embargo, bien puede ser idéntico a otros tipos de estados psicológicos.

El ejemplo del dolor y la ira aducido por Block y Fodor no deja claro si pretenden referirse a la clase de tipos de conductas (o estados neurológicos) que han expresado uno u otro estado psicológico en distintos organismos *de distinta especie* o de la misma. Sin embargo, es difícil dar sentido a la puntualización que sigue salvo que interpretemos que es la variabilidad de instanciaciones físicas (o conductuales) de un tipo de estado mental en distintos organismos *de la misma especie* lo que implícitamente se está discutiendo:

> […O]ne should bear in mind that practically any behavior might, in the appropiate circumstances, become the conventional expression of practically any psychological state and that a given organism in a given psychological state might exhibit almost any behavioral disposition depending on its beliefs and preferences. (Block y Fodor 1972a: 47)

La razón –claro está– es que es difícil dar sentido a la idea de las diferencias en la expresión conductual de estados mentales del mismo tipo venga mediada por distintas *convenciones* excepto en el caso de *Homo sapiens*: han de ser, por tanto, diferencias en la instanciación conductual (o neurológica) de un mismo tipo de estado mental entre especímenes de *Homo sapiens* a lo que Block y Fodor están aludiendo.

Otro de los contratiempos que Block y Fodor (1972a: 48) auguran a quien abrace una doctrina reduccionista de corte disyuntivo incide también en los laberintos a los que conduce la postulación de tesis de identidad cuyos términos sean disyunciones. Es razonable pensar que suceda a menudo que conductas (o estados neurofisiológicos) de cierto tipo se den unas veces como expresión (o sustrato) de estados psicológicos de un tipo, y otras veces de estados psicológicos de tipo diferente. En tales casos, el tipo de conducta (o estado neurofisiológico) en cuestión aparecerá, evidentemente, en las disyunciones con las que identifiquemos cada uno de los estados psicológicos. Pero si cada uno de los términos de esas disyunciones constituye –como se ha planteado– una condición suficiente para albergar el estado psicológico con el cual las identificamos, entonces *todo* sistema que exhiba una

conducta (o un estado neurofisiológico) de ese tipo estará en *ambos* estados psicológicos, lo cual contradice la asunción inicial de que ese tipo de conducta (o estado neurofisiológico) *unas veces* expresa un estado psicológico y *otras veces* otro. De hecho, lo insostenible de la reconstrucción disyuntiva se hace patente en el vocabulario que nos vemos forzados a adoptar para plantearla. En efecto, hemos de describir las conductas como *expresiones* de estados psicológicos, y los estados neurofisiológicos como sus *sustratos*, cuando, en coherencia con el talante reduccionista de las variedades de conductismo y fisicalismo sopesadas, deberíamos hablar ya de conductas, ya de estados neurofisiológicos, como (*idénticos a*) estados psicológicos. La razón de que tengamos que plegarnos a esos giros –digamos– paramentalistas es que resultaría *prima facie* desconcertante –cuando no abiertamente absurdo– plantear *en el seno de una posición reduccionista* que un tipo de conducta (o estado neurofisiológico) es unas veces (idéntico a) un tipo de estado mental y otras veces (idéntico a) otro.

Una vez más, la fuerza de la crítica parece descansar en gran medida sobre la asunción sobrentendida de que la identidad psicofísica (o psicoconductual) de tipos pueda romperse en el mismo o diversos individuos de la misma especie. De no ser así, la acometida de Block y Fodor quedaría sumamente debilitada. El problema se rebajaría –pongamos por caso– al de aclarar lo siguiente: si un organismo de la especie *A* muestra un patrón de activación límbica del tipo L_1, que para su especie constituye estados psicológicos del tipo P_1 y para la especie *B* estados psicológicos del tipo P_2 (P_1 viene dado en la especie *B* –convengamos– por estados neurofisiológicos de tipo L_2), el organismo no se encuentra en un estado psicológico del tipo P_1 *y* en un estado psicológico de tipo P_2. El trámite obvio a tal efecto es incrustar en cada término de la disyunción su rango de aplicación, convirtiendo así el enunciado en una disyunción de conjunciones: albergar un estado psicológico del tipo P_1 equivaldría entonces a ser un miembro de la especie *A* y estar en un estado neurofisiológico de tipo L_1, o ser un miembro de la especie *B* y estar en el estado neurofisiológico del tipo L_2, etc. Pero resulta claro que la dificultad que Block y Fodor tratan de cargar sobre los hombros de sus adversarios filosóficos es más onerosa que ésta.

El golpe de gracia asestado por Block y Fodor a los intentos de usar la noción de identidad disyuntiva para salvaguardar al conductismo o al fisicalismo del embate funcionalista parte de conceder que cupiera, pese a todo, formular una disyunción de (tipos de) conductas (o estados neurofisiológicos) distintiva de cada (tipo de) estado psicológico. Pero aun en tan improbable circunstancia –reza la objeción– nada hace pensar que la correspondencia entre (tipos de) estados psicológicos y (tipos de) conductas o estados neurofisiológicos resultara ser legaliforme. Antes al contrario, cuanto sabemos acerca de la relación entre mente, cerebro y conducta apunta a que lo que encontraríamos sería correspondencias a todas luces *accidentales*. Desde luego, eso es lo que acontecería –como Block y Fodor (1972a: 48) señalan implacables– si a fin de esquivar las anteriores contrariedades

recurriésemos a acotar los términos de la disyunciones conductuales o neurofisiológicas con referencias espaciotemporales: evitaríamos, sí, la profusión de identidades indeseadas, pero a costa de purgar nuestra explicación de toda generalidad y, *a fortiori*, de todo carácter legaliforme.

Proteo también encadenado: nuevos esfuerzos por la unificación de la ciencia

La versión conceptual de la interpretación de la tesis de realizabilidad múltiple como argumento antifisicalista, a la que se aludía *supra* en contraposición a su formulación como hipótesis empírica, se suscita y cobra toda su fuerza, en realidad, en los trabajos de Jaegwon Kim (1989, 1992, 1993, 1998), si bien sólo en el seno del esfuerzo de Kim por desbaratarla. También Wilson y Craver (2007) citan el embate de Kim contra lo que el bautizó como "el mito del materialismo no reduccionista" como una de las dos alas que han hecho alzar el vuelo de la crítica contra las versiones del funcionalismo más despegadas de la tesis de identidad psicofísica y, con ello, contra la noción de realizabilidad múltiple; la segunda, que concerniría a la veta empírica del argumento de Putnam, sería el fulgurante auge de la investigación en neurociencia cognitiva a lo largo de las últimas décadas del s. XX. El ideal de la unificación de la ciencia, que en el nadir del positivismo había tomado los contornos de una quimera, comenzaba a recobrar así parte de su vigor; la naturaleza proteica que habíamos descubierto en lo mental, al mismo tiempo, parecía ir quedando poco a poco subyugada.

El meollo de los argumentos que Kim propone anida en el principio de herencia causal que a su entender atañería a la relación de realización –o "instanciación", según prefiere decir Kim–, por ejemplo, tal como se da entre un estado mental, *M*, y uno físico, *P*:

> […I]f *M* is instantiated on a given occasion by being realized by *P*, then the causal powers of *this instance of M* are identical with (perhaps, a subset of) the causal powers of *P*. (Kim 1993a: 355)

El principio blandido por Kim reposa, como Wilson y Craver (2007) han entrevisto, sobre la nuda intuición materialista respecto a la estructura última de la realidad, a la que el teórico funcionalista, mientras quiera mantener incólume su compromiso con el fisicalismo de casos, se vería fuertemente impelido a asentir[186]. En efecto:

> The intuitive idea behind the principle is that instances of higher-order kinds, such as psychological kinds, inherit the causal powers of instances of the lower-order kinds that realize them. The plausibility of this intuition turns on the instances of each of these kinds

[186] Que el principio de herencia causal está hondamente enraizado en el pensamiento fisicalista se hará patente si recordamos la segunda de las dos tesis en las que Carnap (1963: 883) cifraba lo sustancial de tal concepción del mundo: "[…] all laws of nature, including those that apply to organisms, human beings, and human societies, are logical consequences of the physical laws".

> being the very same glob of matter in the world, or close enough. And if the causal powers of instances are inherited from the physical to the mental, then it seems that all of the real causal action is captured by the lower-order description in terms of physical properties and powers. Thus, it is confused to think of mental causation as somehow distinct from or autonomous of physical causation. (Wilson y Craver 2007: 98-99)

No se incurriría aquí en petición de principios en tanto en cuanto la intuición materialista que nutre el argumento es compartida por fisicalistas y funcionalistas *sensu stricto* –es decir, si se prefiere, por reduccionistas y antireduccionistas.

Sin embargo, es menester replicar que el funcionalista no está obligado a tomarse al pie de la letra dicha intuición *si puede contrarrestarla*, y es posible que en el arsenal conceptual del funcionalismo haya munición bastante para hacerlo. Acaso encontremos los aparejos necesarios en la nítida distinción trazada por Burge (1986, *infra*) entre el carácter local de la causalidad y el, por el contrario, no necesariamente local de la individuación, distinción de la que él se valía para refutar el solipsismo metodológico y que resuena en la defensa del funcionalismo que ensayaría Block (1997) al cuestionar la premisa, asumida por Kim, de que las instancias de una clase natural están ligadas por una relación de proyectabilidad, o quizá en los argumentos con que Davidson (1993, *infra*) intenta resguardar su monismo anómalo del torrente de reparos que el propio Kim le opondría, o en la distinción entre descripción y explicación sobre la que Pylyshyn (1984, *infra*) hacía ya bascular su defensa de la autonomía de la explicación psicológica, imprimiendo sobre la noción de causalidad la distinción entre instancias y tipos en cuyo seno se originara la concepción funcionalista de lo mental[187]. Queda, pues, pendiente la cuestión de si el principio de herencia causal de Kim puede contravenirse, o acaso sólo apaciguarse, sin vulneración del materialismo. En otras palabras: la tesis de que la especificación de propiedades de segundo orden, como serían las psicológicas, no viene acotada en cuanto a su implantación física (Rabossi 1995, *supra*) no es refutada *ipso facto* por la de que dichas propiedades heredan todo su perfil causal de las propiedades físicas en las que quedan implantadas (Kim 1993a, *supra*).

Sea como sea, las objeciones conceptuales de Kim a la tesis de realizabilidad variable componen el substrato en el que se enraíza su severa reprobación de toda lectura antirreduccionista del funcionalismo. Para fijar las premisas fundamentales del argumento es preciso atender a las relaciones entre una extensa familia de conceptos –clase natural, causalidad, proyectabilidad, identidad o coextensividad nómica[188]– que se repite incesantemente, aunque no siempre con plena transparencia, en las disputas contemporáneas sobre la organicidad de lo mental. A juicio de Kim, las clases naturales –o al menos las categorías científicas, que Kim, en un gesto de

[187] *Cf*. Pylyshyn (1984: 11, *infra*).

[188] Debemos a von Wright (1971: 21, nota 62) el recordatorio de que el uso contemporáneo de "nómico" se inaugura en W. E. Johnson (1921/1924, I: *ix*, 7), quien proponía emplear dicho vocablo para describir proposiciones que expresen "una ley natural pura" –lo cual no deja de ser curioso habida cuenta de la etimología de *nomos* y *physis*, y del papel de ambas nociones en el pensamiento de Protágoras o Antifón–, y lo consideraba reemplazable por "necesario" en oposición a "contingente".

realismo científico que Putnam (1981, 1983a, 1998) censuraría, identifica con las clases naturales– agrupan elementos en virtud de la identidad de sus "poderes causales", y cuando una clase de fenómenos queda instanciada, o realizada, en otra, los poderes causales de cada elemento de la clase realizada son idénticos a los de (los elementos de) la clase en la que se realiza: éste es el principio de herencia causal que Kim parece considerar irrenunciable. Pero por otra parte, cabe entender las clases naturales como conjuntos definidos por medio de propiedades proyectables –es decir, según el análisis clásico de Goodman (1953), propiedades cuya verificación en un objeto justifica en alguna medida la expectativa de que otros objetos de la misma clase mostrarán también la propiedad en cuestión. Se sigue de ello, piensa Kim, que cuando dos propiedades *A* y *B* son nomológicamente coextensivas –*i.e.*: cuando la coextensividad de las clases que definen es "al menos nomológicamente necesaria" (Kim 1992: 90)–, *A* será una propiedad proyectable si y sólo si *B* lo es también. El dilema conceptual en que cobra cuerpo la posición de Kim parte de conceder al funcionalista que un tipo de estado mental *M* venga encarnado en una disyunción heterogénea $F_1 \vee F_2 \ldots \vee F_n$ de tipos de estados físicos, y adopta la siguiente forma: dado que, *ex hypothesi*, la propiedad de atravesar un estado mental del tipo *M* y la de atravesar un estado físico del tipo $F_1 \vee F_2 \ldots \vee F_n$ definen clases nómologicamente coextensivas, o bien tanto "atravesar un estado mental del tipo *M*" como "atravesar un estado físico del tipo $F_1 \vee F_2 \ldots \vee F_n$" son propiedades proyectables, o bien ni una ni otra lo son; luego o tanto *M* como $F_1 \vee F_2 \ldots \vee F_n$ son clases naturales, o ni una ni otra lo son. Ahora bien, si tanto *M* como $F_1 \vee F_2 \ldots \vee F_n$ son clases naturales, no hay razón para poner en duda la reductibilidad de *M* a $F_1 \vee F_2 \ldots \vee F_n$, como pretendían Putnam o Fodor; si, en cambio ni una ni otra son clases naturales, entonces no habrá una ciencia natural de $F_1 \vee F_2 \ldots \vee F_n$ –no cabe una neurofisiología del dolor, o de la alegría...– pero tampoco podrá haberla de *M* –no hay lugar para una psicología de esos tipos de estados mentales. El primer desenlace es una rotunda reivindicación del reduccionismo. El segundo –como ha quedado recalcado– galantea con el eliminacionismo, del que sólo se aparta tímidamente por la convicción de Kim de que es viable desarrollar teorías científicas –psicológicas, fisiológicas, químicas, físicas: cada una reducible a su sucesora– para clases adecuadamente restringidas, como pudiera ser determinada variedad de dolor en humanos –psicologías provincianas y sufragáneas, decíamos.

Pero el argumento conceptual de Kim embarranca si se exige una aclaración del significado preciso que se quiere otorgar a la expresión "nomológicamente coextensivas", referida ya a pares de propiedades, ya a pares de clases –o, en general, *n*-plas de propiedades o de las clases que ellas definen–, y al concepto de proyectabilidad. Porque si se entiende, al pie de la letra, que dos clases son nomológicamente coextensivas cuando ambas contienen los mismos elementos y ello es el resultado necesario de leyes naturales, está a la mano del funcionalista negar que tal relación de coextensividad entre dos clases entrañe que la proyectabilidad de la propiedad que define una de ellas coimplique la proyectabilidad de la propiedad que define la otra. Se trataría, entonces, de hacer patente la posibilidad de que dos

propiedades *M* y *F* sean, o definan clases, nomológicamente coextensivas en el sentido especificado, y de que, aún así, del hecho de que *MxFx* no se siga que la predicción de *Fy* esté justificada por la constatación de *My*. En efecto, Block (1997) ha mostrado que la noción de proyectabilidad que Kim necesita para sostener su argumento no es la habitual, relacionada con un criterio epistemológica de *justificación* de una predicción, sino una versión fortalecida de ésta, en la que el papel decisivo lo desempeñaría cierta idea de *evidencia objetiva*, ajena a las vicisitudes de la justificación, que depende de la evidencia disponible para un sujeto. Así, resulta que:

> […t]he usual notion of projectibility is not what Kim needs, for on the usual notion having to do with justified belief, we could be justified in supposing that something will have one of two nomically equivalent properties without being justified in supposing it will have the other. (Block 1997: 116)

O, en menos palabras:

> The usual notion of projectability may be too epistemic to bear the metaphysical weight that the notion of a kind is supposed to bear. (Block 1997: 116)

Una vez provista de esa noción de proyectabilidad objetiva, la argumentación de Kim parece tornarse inexpugnable. Tanto –es preciso anotar– que descuida uno de los primeros flancos abiertos en la crítica del reduccionismo psicofísico, logrando una aparente invulnerabilidad en un terreno a costa de dejar otro prácticamente desguarnecido. Así se pone de manifiesto si atendemos al análisis desplegado en Boyd (1980) del compromiso de las primeras defensas de la tesis de identidad psicofísica –las de Place (1956) o Smart (1959), por ejemplo– con una concepción lockeana, aunque heredada por vía de Hume, de la naturaleza de las clases naturales –las "especies distintas" (Locke 1690: *passim*) del *Ensayo sobre el entendimiento humano*. Es sabido que para Locke –y desde entonces, para buena parte de la tradición empirista– las clases naturales carecen de otra esencia que su esencia nominal: las propiedades esenciales del agua o el dolor lo son únicamente en el sentido y en la medida en que sean parte del significado de "agua" o "dolor", significado que viene delimitado convencionalmente. El convencimiento de que la experiencia puntea los límites del conocimiento lleva a Locke a negar que respecto a la esencia real de las clases naturales pueda darse otra cosa que baldía especulación metafísica. Cuando las primeras formulaciones de la tesis de identidad psicofísica eran denunciadas por su presunta violación de la indiscernibilidad de los idénticos –pues si un dolor, digamos, punzante fuera idéntico a cierto patrón de activación nerviosa esa faceta de la ley de Leibniz nos forzaría a considerar punzante también la activación nerviosa, conclusión a la que tanto los reduccionistas como el sentido común se resisten–, Place, Smart o, más abiertamente, Feigl (1958, *supra*) se defendían apelando a que un único y mismo estado, al que se refieren tanto la descripción psicológica como la fisiológica, es mencionado en aquella descripción por una propiedad introspectiva, cual es su carácter punzante, y en esta otra descripción por determinadas

propiedades fisiológicas: subyacería a las críticas, así pues, cierta ceguera ante la distinción fregeana entre referencia y sentido. Sin embargo, la objeción no alude a las propiedades por las que el estado en cuestión sea mencionado, sino a las que de hecho posee: lo que se señala es la obligación en que incurre el reduccionista de aceptar que la activación nerviosa *es* punzante, so pena de vulnerar la ley de Leibniz, no la de aceptar que al describir tal estado como activación nerviosa mencionemos su carácter punzante. Así que para perseverar en su defensa, el reduccionista se ve impelido a dar el paso de afirmar que el estado en cuestión *posee de hecho* la propiedad de ser punzante *bajo la descripción psicológica* pero no la posee de hecho bajo la descripción fisiológica. Ahora bien, ese paso –como señala Boyd– sólo es practicable bajo una concepción lockeana de las clases naturales, en la que cuáles sean las propiedades esenciales de una clase natural es asunto relativo a cómo decidamos, convencionalmente, delimitarla. Enmarcada en esa concepción lockeana, la réplica reduccionista aparece entonces más transparente: es parte del significado de "dolor" que el dolor tiene ciertas propiedades introspectivas, como su carácter punzante, pero tal cosa no es parte del significado de "patrón de activación nerviosa de tipo *C*". O, más en general:

> Contingent identity statements entail that the identified entities have the same properties, but (since essential properties are description-dependent) they do not entail that the identified entities have the same essential properties. (Boyd 1980: 69)

La clave, en efecto, del alegato reduccionista reside en la afirmación de que la tesis de identidad psicofísica declara una verdad contingente, fruto de descubrimientos empíricos: si se tratara de una verdad necesaria, entonces –al menos bajo el rígido esquema de las relaciones entre necesidad, aprioricidad, analiticidad y sinonimia que escolta a la concepción de las clases naturales heredada de Locke y Hume– el término mentalista y el término fisicalista del enunciado de identidad serían estrictos sinónimos, de modo que no ofrecerían diferentes descripciones del fenómeno a las que cupiera apelar. Pero –asegura el reduccionista– no es el caso.

La sagaz observación de Boyd, dicho sea de paso, es que la propia concepción lockeana de las "especies distintas" que se emplea en su defensa bloquea la identificación de clases de estados mentales con clases de estados físicos. La pretensión –esto es– de identificar, en virtud de haber descubierto su correferencialidad, dos clases cuyas esencias nominales son diferentes no sólo quiebra el principio de que las clases naturales vienen fijadas únicamente por sus esencias nominales –de que "[…] cada idea abstracta dotada de nombre constituye una especie distinta […]" (Locke 1690: III, VI; §38)–, sino que, por empírico que se diga el descubrimiento de la correferencialidad de ambas, desborda los límites de la experiencia, que son los del conocimiento. Así:

> What is ruled out […] by a Lockean account of general terms –and by the associated empiricist epistemological outlook– is the view that, although we do not classify pains as physical, nevertheless pain poses the same essential features as do paradigmatically

> physical states, and we could eventually discover that they are really physical. According to the Lockean analysis, all there is –or could be– to being physical is having the properties conventionally taken to be marks of the physical. (Boyd 1980: 72)

Desde este punto de vista, la respuesta de Place, Smart o Feigl a las objeciones antirreduccionistas era frágil en cualquier caso. Mientras no renuncie al aliento antimetafísico que da vida a la concepción lockeana de las clases naturales, sólo le quedan dos rutas cabales: convertir su tesis en una propuesta revisionista acerca de las convenciones que hemos de adoptar en cuanto al significado del vocabulario mentalista, o, de nuevo, abandonarse a la eliminación de esos términos. De acuerdo con la recapitulación elaborada por Boyd (1980: 74), la primera habría sido la trayectoria elegida por Shaffer (1961) –proponer una modificación de nuestra convenciones lingüísticas de modo que quepa atribuir a estados mentales predicados que hasta ahora considerábamos adecuados únicamente para estados físicos, y viceversa– o Feyerabend (1963, pero *cf. supra*). La vía de la eliminación, en cambio, habría sido explorada por Rorty (1965), en la estela de Quine (1960), al reconstruir la noción de identidad empleada en el programa reduccionista en términos de desaparición de la entidad inexistente –el (tipo de) estado mental– cuya identidad con una entidad existente –el (tipo de) estado físico– se formula. Así visto, el eliminacionismo aparece también, al fin y al cabo, como una iniciativa de revisionismo semántico, aunque más drástica. Como anota Boyd, lo medular de la moción expuesta por Rorty es la idea de que:

> The statement "My thought at *t* = brain state *B*" really says that there is no such thing as my thought that *t* but that brain state *B* is what we should talk about instead. Since there are no thoughts –and hence no nagging thoughts– the problem of predicating naggingness of brain states does not arise. And, similarly, for other difficult cases of mind-brain identity. (Boyd 1980: 74)

Por lo demás, toda la constelación de trabajos en torno al concepto de identidad teórica que acompaña al auge de la tesis de identidad metafísica (Cornman 1962, Nagel 1965) e incluso –cabe agregar– todo el desarrollo del análisis temáticamente neutral por parte de Smart (1959, *infra*) obedecen a un intento –malogrado, a juicio de Boyd (1980)– de aliviar la tensión entre la tesis de identidad psicofísica y la visión lockeana de las "especies distintas", sin rendir esta última. Pero eso es –si la razón asiste a Boyd– justo lo que el materialismo debe hacer: deponer la animosidad antiesencialista de las reflexiones de Locke sobre las clases naturales, y proveerse de un modo más robusto de entenderlas[189].

[189] Una alternativa a la concepción lockeana de las clases naturales es la desarrollada por Kripke (1972/1980) en el marco de su análisis de las nociones de necesidad y referencia. Las severas críticas que el propio Kripke empuña contra toda forma de materialismo –inspiradas en lo que él mismo denomina "intuiciones cartesianas"– parecen desaconsejar la adopción de su propuesta si el objetivo es amparar la tesis de identidad psicofísica. No obstante, Boyd considera que la crítica del materialismo de Kripke es inocua, mientras que la fundamentación del materialismo en una

Aunque cuando al renunciar a la concepción lockeana de las clases naturales en beneficio de una en la que éstas vienen delimitadas por su proyectabilidad objetiva, o por los poderes causales compartidos por sus elementos, Kim pareciera haber escuchado la arenga de Boyd (1980, *supra*), al hacerlo queda inerme ante la pregunta de si el carácter punzante con que algunos dolores nos atormentan es una propiedad de una cierta clase –acaso disyuntiva– de patrones de activación fisiológica, y así, una vez más, abocado a seguir los pasos de Quine o Rorty. Desde luego, que un dolor sea punzante no es comprensible si no posee la capacidad de infligirnos tales punzadas, y eso parece un "poder causal" como el que más: éstos –dice Kim– los hereda la propiedad reducida –"sentir dolor", supongamos– de la reductora –"atravesar un patrón de activación fisiológica $F_1 \vee F_2 \ldots \vee F_n$", así que deben contarse entre las propiedades de cada término de la disyunción. Pero si es así, no había ninguna necesidad de delimitar esa clase mediante un acúmulo de disyunciones: podíamos hacerlo sencillamente apelando a esa propiedad común. Lo que parece estar ocurriendo, en suma, es que, a fin de evitar que la heterogeneidad de los tipos físicos en los que pueda encarnarse un tipo mental sirva de sustento a la autonomía de la explicación psicológica, las estipulaciones de Kim respecto a los lazos entre los conceptos de clase natural, eficacia causal, proyectabilidad y coextensividad nomológica estrechan tanto su noción de identidad de tipos que la vuelven a duras penas sostenible. Dicho de otro modo: Kim dicta la identidad de todas y cada una de las propiedades causalmente eficaces de los elementos de aquellas clases que empíricamente resulten ser, como resultado de leyes naturales, coextensivas, con completa independencia de los criterios que hayan servido para delimitar las clases. Esto cercena la posibilidad de que unas u otras de dichas clases intervengan en explicaciones autónomas, pero sólo a fuerza de ignorar la de que cada una de ellas –es decir, sus elementos– posean *otras* propiedades causalmente eficaces –o cuando menos explicativamente relevantes en algún sentido– distintas de las que comparten entre sí[190].

De índole muy diferente son las objeciones al argumento antifisicalista de Putnam que tempranamente urdiera Richardson (1979), y que conviene dejar anotadas aunque sea al vuelo. A juicio de Richardson, los planteamientos de Putnam dañarían todo lo más a un proyecto de reducción psicofísica que pretendiera fundamentarse en una fantasmagórica concepción de la reducción interteórica que a duras penas se reconoce en la articulada por Ernest Nagel (1961). La razón es que el principio de derivabilidad que, según Nagel, se cumple entre una teoría reducida y su teoría reductora –en virtud del cual aquélla es derivable de ésta– no exige más que enunciados condicionales que expresen, en el vocabulario de la teoría reductora, las condiciones suficientes para que se verifiquen los enunciados de la teoría reducida. Si

concepción de las clases naturales capaz de evitar las trabas que ofrece la de Locke es imperiosa. Sobre las nociones de esencia y clase natural en Locke y Kripke, *cf*. Mackie 1976: 72-106

[190] En esta misma línea puede entenderse tal vez la réplica de Davidson (1993, *infra*) al esforzado desmantelamiento de su monismo anómalo por parte de Kim (1992); tácticas afines –como ya se ha adelantado– veremos entre otros en Pylyshyn (1984) y Burge (1986).

bien la noción de "leyes-puente" –ajena al glosario de Nagel (1961), que se expresa en términos de "condiciones de conectividad" (Nagel 1961: 543)– ha terminado por heredar el carácter bicondicional de las equivalencias materiales que solían aparecer en los ejemplos históricos aducidos por el propio Nagel, lo cierto es que para reducir la psicología a la física, según el modelo nageliano de reducción interteórica, lo único que hace falta es dar con condiciones físicas suficientes para que se registren cualesquiera condiciones psicológicas. Los argumentos de Putnam, así pues, devastarían todo proyecto de reducción psicofísica que requiriese fijar condiciones físicas necesarias –y, además, como apunta Bickle (2006: §2.1) en honor a Lewis (1969), no disyuntivas– para cualesquiera condiciones psicológicas, pero tal proyecto no existe. Mientras el objetivo de nuestas investigaciones sea una reducción según los parámetros descritos por Nagel (1961), los reparos de Putnam ni siquiera nos rozan. La mala noticia, para Richardson, es que los propios defensores de la identidad psicofísica parecen haber malentendido a Nagel tanto como Putnam, y han perseverado en la empresa de rastrear las propiedades físicas $F_{1...n}$ que pudieran quedar ligadas a cada propiedad psicológica $P_{1...n}$ no por una relación consistente en que F_i instaure condiciones suficientes para P_i, y nada más, sino por una relación de identidad –aun a fuerza de admitir, como Lewis (1969) o Kim (1989, 1992), que las propiedades físicas en cuestión hubieran de quedar delimitidas mediante disyunciones, o bien, como Kim (1998), de intentar forjar un modelo de reducción interteórica al margen del erigido por Nagel[191]. Es indudable, en todo caso, que tanta disconformidad en la exégesis de Nagel no es azarosa, sino que apunta a una ambición epistemológica que, sin quedar recogida en el concepto de reducción interteórica apuntalado por Nagel, nos resulta difícilmente declinable. La misma ambición, por cierto, que Block (1978, *infra*) esgrimiera contra Lewis (1969).

Frente a la versión empírica del argumento de realizabilidad múltiple, en cambio, Rabossi (1995) –siguiendo de cerca los pasos de Lewis (1969) y Kim (1972, 1982, 1989), acaso también los de Lycan (1981)– se inclina por la estrategia de circunscribir el ámbito del proyecto reduccionista de la tesis de identidad a los confines de cada especie, marcando como objetivo explicativo enunciados como "Los miembros de la especie *E* tienen [un estado psicológico de tipo] *x* cuando están en […] [un] estado cerebral [de tipo] *c*" (Rabossi 1995: 38). La impugnación de esta estrategia por parte de Block y Fodor (1972a), por lo demás, parece antojársele completamente inane: que adoptar un reduccionismo de envergadura restringida nos deje sin conceptos universales de cada tipo de estado mental es, en un sentido, irrelevante –¿por qué habríamos de tener tales conceptos?–, y, en otro, falso –gracias precisamente a que *sí* poseemos tales conceptos, articulamos los distintos enunciados de identidad psicofísica como enunciados acerca del mismo tipo de estado mental en distintas especies. Pero tanto lo desdeñoso del primer proceder esbozado como lo

[191] Hay una templada evaluación de ese intento en Marras (2006), quien concluye que la diferencia entre identidad de una propiedad reducida con la reductora y explicación de una propiedad reducida mediante la reductora persiste tanto en el modelo de Nagel como en el de Kim, y es suficiente para respaldar las reivindicaciones antifisicalistas.

estipulativo del segundo resultan poco convincentes: poseer tales conceptos sería interesante exactamente para hacer con ellos lo que en el segundo intento de esquivar la cuestión se dice que hacemos, pero si efectivamente podemos hacer tal cosa merced a conceptos generales de tipos de estados mentales que sí poseemos, entonces hace falta argumentar por qué esos conceptos generales no merecen ser identificados con el tipo de estado mental que denotan, y sí lo merecen los conceptos fisicalistas de ámbito más reducido y tono más reduccionista.

El debate sobre la viabilidad de estas –digamos– reducciones interteóricas acotadas, o –como prefería plantearlo Lewis (1969)– relativas al contexto, ha orbitado a menudo en torno al caso de la temperatura o, más ocasionalmente, de la hidrosolubilidad, o la acidez y la alcalinidad. La defensa del fisicalismo de tipos inspirada en la historia de la termodinámica ha quedado sucintamente compendiada por Bickle (2006: §2.2) al hilo de los trabajos de Hooker (1981), Enç (1983) y P.S. Churchland (1986); la réplica funcionalista se veía ya anticipada en Block (1978: 56, *supra*). La propiedad denominada "temperatura" en el ámbito de la termodinámica clásica se reduce de hecho a diferentes propiedades físicas para diferentes estados de la materia, pero ello no es óbice –reza el argumento fisicalista– para que registremos el conjunto como un caso paradigmático de reducción interteórica, y no hay motivo para ser más exigentes cuando se trata de la reducción de propiedades psicológicas. No se trata ya, como había adelantado Block (1978) de que innumerables objetos tan diversos en su estructura física como podamos imaginar puedan exhibir una temperatura, hecho que Block daba por obvio y resolvía apelando a que la misma propiedad física –energía cinética media– constituía la temperatura en todos los casos. Lo que ahora se plantea es que distintas propiedades físicas vendrían a constituir la temperatura según el objeto de marras fuese un sólido o un líquido, un gas –peor aún: según si fuera un gas monoatómico, diatómico, o multiatómico– o un plasma, o incluso el vacío. La cuestión, entonces, es si una somera indagación del concepto termodinámico de temperatura devuelve la razón a Block (1978) en cuanto a que pueda concebirse a su vez una propiedad reductora para todas esas diversas propiedades reductoras de la temperatura. Dicho de otro modo, ¿es radicalmente heterogéneo –como hacen ver los argumentos fisicalistas– el conjunto de propiedades que identificamos con la temperatura en distintos estados de la materia, o cabe intuir en él algún principio homogeneizador? Pues bien, no parece particularmente aventurado afirmar que la referencia a los grados de libertad en la trayectoria –de traslación, rotación o vibración– de las partículas que forman el cuerpo físico cuya temperatura nos interesa, así como a la absorción y emisión de radiación electromagnética –decisivas cuando no existen tales partículas, es decir, en el vacío–, apuntan precisamente a una unificación de las propiedades reductoras de la temperatura en torno a la noción de "energía sensible", a la que se adjudicaría el papel, como propiedad de segundo orden, de enlazar la propiedad reducida –la temperatura–, a la que es idéntica, con la propiedad reductora unificada. Estimar la plausibilidad de esta impresión, por supuesto, desborda con mucho el alcance de este trabajo; es de rigor, pues, contentarse con no dar por concluyentes los argumentos

fisicalistas apoyados en la interpretación de la termodinámica. De todas maneras, el esfuerzo sería en buena medida vano, puesto que aun si la heterogeneidad de las propiedades físicas que subyacen a la temperatura en distintos estados de la materia resulta ser irreconciliable, entonces o el concepto de temperatura acreditará su relevancia explicativa ofreciendo con su concurso la posibilidad de capturar generalizaciones que se nos escaparían si restringiésemos nuestro vocabulario teórico al de esas heterogéneas propiedades en que la temperatura puede quedar instanciada –posibilidad que Kim (1993a: 355, *infra*), en el caso de los estados mentales, considera anulada por un principio de herencia causal–, o no lo hará. Si lo hace, estará en manos del funcionalista concluir que el de temperatura es de hecho un concepto funcional –como venía advirtiendo desde el principio–; de lo contrario, se impondrá por fin el ultimátum eliminacionista. De cualquier forma, la apelación a que la reducción de la temperatura a dichas propiedades sea un ejemplo paradigmático de reducción interteórica no pasaría de ser un argumento *ad auctoritatem*, al que bien cabría responder describiendo ese hecho como manifestación de un injustificado sesgo reduccionista y reclamando ya la reconstrucción funcionalista del concepto de temperatura ya su rigurosa eliminación.

Ante un embate eliminacionista de ese calado, de llegar a darse, sólo quedaría al reduccionista[192] el trámite ensayado por Lewis (1969: 233, *infra*) en relación con la identidad psicofísica: denunciar que la repulsa de los conceptos cuya referencia en distintos contextos sea radicalmente heterogénea es un prejuicio anticuado. La legitimidad del afán que Lewis desprecia descansa sobre la convicción de que nuestros conceptos mentalistas ordinarios, o al menos los referidos a tipos de estados psicológicos, son aplicables con propiedad a organismos de distintas especies, y podrían serlo incluso a sistemas computacionales o a organismos que no estuvieran basados en la química del carbono; es decir, la convicción de que las propiedades que esos conceptos aíslan son transversales a la distinción entre especies, entre sistemas cognitivos vivos y otros sistemas de procesamiento de información, etc. –esa es la idea que sustenta la propuesta de Miller (1984 *apud* Pylyshyn 1984: xiii, *supra*) de acuñar el neologismo "informávoros" para referirnos a los sistemas que muestran dichas propiedades. Lo llamativo de la posición de Lewis (1969) es que parece compartir esta convicción, pero no se muestra dispuesto a dar legitimidad a la ambición de desplegar conceptos de estados mentales que no sean dependientes del tipo de sistema al que se atribuyan. O, si no quisiéramos adentrarnos en el camino abierto por Lewis, podríamos quizá, como por momentos parece intentar Kim (1992), remedar el quiebro que Quine (1960: 280, *supra*) intenta dar a toda defensa de la autonomía explicativa de la psicología que se fundamente sobre variantes de la idea de que las propiedades que articulan los tipos de estados que ésta describe no son coextensivas con propiedades físicas: convertir las premisas de dicha defensa en un síntoma de la impotencia explicativa de los constructos psicológicos, *ergo* de su

[192] Naturalmente, para quien abandone el compromiso con el reduccionismo, como Putnam (1988), hay otros caminos abiertos, que el propio Putnam ha tratado de tantear.

inutilidad. Pero, por supuesto, seguir los pasos de Quine en este punto es abrazar el eliminacionismo, y eso no termina de compadecerse bien con el espíritu de otros muchos argumentos de Kim. Si bien lo único que Kim se muestra dispuesto a eliminar son los referentes de los conceptos que describen tipos interespecíficos de estados mentales –es decir, que lo que se nos propone es sólo renunciar al proyecto de una psicología científica válida para organismos de distintas especies o sistemas de otra naturaleza, no a todo proyecto de psicología científica–, sabemos ya que el impulso de la tesis de realizabilidad variable no se agota ahí. Si hubiera razones para creer que estados mentales de un mismo tipo pudieran desgranarse en estados neurofisiológicos de distinto tipo en distintos individuos, o en distintos momentos de la vida de un único individuo, la comedida renuncia de Kim habría de ceder a un eliminacionismo a secas. Según anota Bickle (2006: §2.6), circunspecto pero inadvertido del reto eliminacionista:

> The more radical type of multiple realizability seems to force increasingly narrower domains for reductions to be relativized –at the extreme, to individuals at times. This much "local reduction" seems inconsistent with the assumed generality of science.

Menos aún, por cierto, parece reparar Block (1997) en los tintes eliminacionistas que impregnan el razonamiento de Kim. La cuestión, para Block, es que Kim ha obligado al antirreduccionista a afrontar un dilema: o tanto los estados psicológicos como las disyunciones heterogéneas a las que estos son nomológicamente equivalentes son clases naturales (que es tanto como decir, a juicio de Kim, que constituyen propiedades proyectables), o no lo son ni unos ni otras. En el primer caso, quedaría avalada la reducción de cada tipo de estado psicológico a un tipo de estado físico, aunque se trate de un tipo disyuntivo, de modo que el antirreduccionismo caería en desgracia. En el segundo caso, los estados psicológicos no son clases naturales, y por tanto no es viable una ciencia que los tome como objeto de estudio. El dictamen de Block es tan rotundo como provisional: "[…] comoquiera que sea, el reduccionismo derrota al argumento de realizabilidad múltiple" (Block 1997: 114). Pero el propio razonamiento de Block lo aboca rápidamente al mismo atolladero en que se viera Bickle:

> […E]ven if there are no *general* psychological kinds, there can nonetheless be restricted psychological kinds that are not multiply realizable with respect to lower level science. Pain in general and thought in general are multiply realizable. If that makes them non-kinds, perhaps human pain or human thought is not multiply realizable, or if not, Ned-pain or Ned-thought or Ned-pain-now. And so there is room for these restricted kinds to be reducible to physics and chemistry. (Block 1997: 115)

Es difícil, no obstante, dar sentido a la idea de que pudiéramos desarrollar una ciencia del dolor que siente un sujeto en un momento dado –como Ned, ahora– toda vez que dicha ciencia no nos permitiría generalización alguna a otros sujetos ni otros momentos, lo que, dicho sea de paso, supone que tampoco nos permitiría predicción

alguna. La ciencia de los estados psicológicos sería entonces una ciencia puramente idiográfica[193], y el logro de su reducción al vocabulario teórico de disciplinas más básicas una victoria pírrica en la medida en que dichas disciplinas albergaran alguna ambición de ofrecer explicaciones nomotéticas. Acaso el reduccionista *tenga la suerte* de que la realizabilidad múltiple de los estados psicológicos se detenga de hecho antes de llegar al nivel del individuo y el instante, o acaso sea factible construir un argumento más o menos conceptual que nos persuada de que tal cosa ha de suceder, pero no parece que haya nada parecido en las reflexiones de Kim. En todo caso, si no son clases naturales ni estados psicológicos como el dolor o el pensamiento ni las disyunciones de estados físicos en que estos se encarnan, entonces –cierto– no habrá una psicología del dolor o del pensamiento, pero tampoco habrá una neurología, una bioquímica ni una física de la encarnación neurológica, bioquímica o física del dolor ni del pensamiento. O, como dice Block (1997: 115), "[…] las únicas […] ciencias especiales que existen son aquellas que resultan reducibles a la física y la química […]" –que es tanto como decir que *no hay ciencias especiales*, dado que por "especiales" solemos entender precisamente "irreductibles". Esto es sin duda una derrota para el funcionalismo, y también una conquista para el eliminacionismo. Su reivindicación como una victoria del reduccionismo, en cambio, es poco creíble, salvo en el paupérrimo sentido en que toda derrota del antirreducionismo es una victoria del reduccionismo. Ahora bien: el legítimo empeño fundacional de la concepción reduccionista de la ciencia no era reducir las ciencias especiales a ciencias básicas *después de haberlas desprovisto de contenido* –o, al menos, de contenido nomotético. En suma, reivindicar el reduccionismo exige tanto escrúpulo en desarbolar el funcionalismo como en esquivar el eliminacionismo, y parece claro que Kim ha puesto más ahínco en lo primero –aún hemos de ver si provechosamente.

Una expresión acaso más tajante de algunos requiebros del pensamiento de Kim es la propuesta ensayada por Zangwill (1992): negar que existan los mismos tipos de estados psicológicos en diferentes especies o, como apunta Bickle (2006: §2.4), que los estados psicológicos de diferentes especies se inserten en los mismos "[…] patrones de causas y efectos salvo en la descripción más burda". Resulta apenas discutible que aceptar la tesis de Zangwill nos obligaría a renunciar a la de realizabilidad variable. En efecto, si no existen tipos psicológicos interespecíficos, no tiene más sentido atribuir a dichos tipos una mudable encarnación física que atribuirles las virtudes de la piedra filosofal o del Ave Fénix. No es menos evidente,

[193] La distinción entre un saber nomotético, propio de las ciencias naturales, y uno idiográfico, característico de las ciencias sociales o históricas, que debemos a Windelband (1894), ha sido, junto con la distinción entre explicación y comprensión (*infra*), a la que de hecho se hermana al vincular la ciencia nomotética a procesos de abstracción y la idiográfica a actos de intuición, uno de los pilares de las corrientes antipositivistas que han permeado la filosofía de la ciencia desde la misma génesis del positivismo. No es difícil, como veremos, hallar en el seno del cognitivismo huellas de dichas corrientes, si bien la aspiración a proporcionar explicaciones de índole nomotética –además de descripciones idiográficas cuando sea preciso– parece haber permanecido incólume. Un interesante estudio del modo en que, sin embargo, el sentido de la distinción en su uso psicológico se ha ido apartando de los planteamientos originales de Windelband puede encontrarse en Lamiell (1998).

ahora bien, que aceptar la tesis de Zangwill es tanto como aceptar la falsedad del funcionalismo, y –noticia menos grata al propio Zangwill– que no constituye una reivindicación del fisicalismo de tipos –cuyas premisas desbarata en igual medida que las del funcionalismo–, sino, siquiera débilmente, del eliminacionismo. Si Zangwill está en lo cierto, la identificación del referente de nuestros conceptos mentalistas ordinarios –al menos, los que se comportan de hecho coloquialmente como conceptos interespecíficos– con estados funcionales de los organismos o sistemas que los albergan sólo cuadraría cuando la caracterización de esos estados funcionales fuese irremisiblemente vaga. Pero como esa identificación es precisamente el proyecto funcionalista, y nadie ha dado ese proyecto por cumplido[194], lo más que Zangwill puede pretender sin incurrir en *petitio* es contrarrestar los indicios empíricos de verosimilitud de la tesis de realizabilidad múltiple que, desde Putnam (1967a), suelen darse por sentados en la literatura funcionalista. Eso, por supuesto, no liquida el proyecto funcionalista, sino que forma parte integral de él, en tanto que programa de investigación científica: lo sorprendente sería no encontrar atisbo alguno de observaciones que obligaran a remodelar nuestras hipótesis. Desde luego, ha de formar parte de esa evaluación dirimir si lo que Zangwill tilda de vaguedad irremisible podría verse, con mejores ojos, precisamente como las herramientas de abstracción necesarias para desarrollar conceptos psicológicos interespecíficos. En suma, los argumentos de Zangwill son sumamente valiosos de cara a la evaluación empírica de las tesis funcionalistas, pero aceptar su conclusión sería precipitar el desenlace de esa evaluación dando por falso el funcionalismo. De cualquier manera, es preciso conceder que la conclusión que encabeza el trabajo de Zangwill –que la tesis de realizabilidad múltiple no está demostrada– es del todo aceptable; de nuevo, es una pena que su antítesis no se cuente entre las reivindicaciones de –como diría Lewis (1969)– ningún funcionalista "razonable".

En cuanto a que la reivindicación del fisicalismo de tipos elude a Zangwill, conviene recordar que la tesis de identidad psicofísica, incluso en su versión de identidad de especificaciones funcionales propugnada por Lewis, depende de la existencia de tipos psicológicos interespecíficos tanto como el propio funcionalismo, puesto que son esos tipos de estados psicológicos interespecíficos los que se hacen equivaler a (una disyunción de) tipos de estados físicos. Además, los argumentos empleados por Zangwill contra la existencia de tipos psicológicos interespecíficos, articulados en torno a las diferencias sensoriales y motoras que distancian a organismos de distintas especies, hacen impensable que pueda emprenderse, desde esa constatación, una defensa de la existencia de tipos físicos interespecíficos, según la cual las descripciones funcionales de –digamos– los estados de dolor serían desiguales en distintas especies, pero sus descripciones neurofisiológicas o físicas fueran después de todo equivalentes. Así que si el desmantelamiento de la noción de tipos psicológicos interespecíficos ensayado por Zangwill es efectivo, el fisicalismo

[194] El propio Putnam (1997: 35, *supra*), por el contrario, se muestra muy severo con las promesas incumplidas del funcionalismo.

de tipos sufre tanto como el funcionalismo. Quien, en cambio, cobra fuerza es el eliminacionista, igual que sucedería si el heterogéneo conjunto de propiedades físicas denotado por el concepto de temperatura resultara ser radicalmente heterogéneo: si el dolor de uno de nosotros y el de uno de los pulpos imaginados por Putnam son estados irreconciliablemente dispares, lo que procede es forjar sendos conceptos para cada uno de ellos y deshacernos de una vez por todas del equívoco concepto de "dolor"[195]. Quedaría por ver, además, si las diferencias entre individuos de una misma especie, o distintos momentos en el mismo individuo, resultan para Zangwill suficientemente leves como para permitir la existencia de tipos psicológicos interindividuales, o incluso intraindividuales, o, por el contrario, suficientemente profundas como para aconsejar su desarticulación, dando pábulo a la voracidad eliminacionista.

La misma pregunta gravita, como se ha visto, sobre los argumentos de Kim (1992), que concede que determinadas clases de estados mentales pudieran verterse sobre clases físicas irreparablemente disyuntivas, debido a su movediza materialización, pero asegura que tal cosa, lejos de ratificar la autonomía de la psicología, muestra que no es posible desarrollar una psicología científica supraespecífica –es decir, que no existe la forma normal de descripción psicológica imaginada por Putnam (1960: 43; 1967a: 227, 229, *infra*). Sólo habrá, pues, psicologías específicas, plenamente reducible cada una de ellas a la neurofisiología del tipo de organismo o sistema del que se trate, y en gran medida aisladas entre sí –salvo por las pasarelas que ocasionalmente puedan alzarse en el marco de la teoría neurofisiológica reductora: psicologías provincianas, heterónomas –sufragáneas de la fisiología nerviosa–, pero no una psicología cosmopolita ni, desde luego, soberana. Ahora bien, la pregunta que, como Zangwill, debe afrontar Kim es cómo de provincianas serán esas psicologías o fisiologías, o por qué hemos de esperar que sus lindes respeten los de cada especie biológica. Ante el horizonte de teorías cuyo ámbito de aplicación pudiera estrecharse hasta abarcar no más allá del individuo, o incluso del individuo en un instante de su biografía, no está claro cuánto tiempo podría prolongar Kim su renuencia al giro quineano de desertar de los conceptos mentalistas.

El dolor y la piedra de ijada: coextensividad nomológica y herencia causal

Hemos recalado en la conclusión, así pues, de que el funcionalista no deviene obligado por el razonamiento de Kim a optar entre un reduccionismo de amplio espectro o uno ceñido a ciertas restricciones sobre nuestros conceptos ordinarios, más afín al espíritu eliminacionista. Conviene constatar, en todo caso, que al desenvolver

[195] De hecho, los reparos de Zangwill respecto a la existencia de tipos psicológicos interespecíficos (definidos en términos funcionales, se entiende) lindan con el convencimiento al que habría de llegar el propio Putnam (1988: *xiv, infra*) de que los tipos de estados mentales son computacionalmente plásticos además de composicionalmente plásticos, y la tarea que con mayor urgencia aborda Putnam tras alcanzar esa convicción es la de silenciar los previsibles vítores eliminacionistas.

sus consideraciones Kim ofrece, a su pesar, motivos de peso para inclinarse por el segundo desenlace. Dichos motivos se destilan en la apelación al caso del jade como contrapunto de la diversa encarnación que pudieran mostrar los estados mentales: como apuntara Putnam (1975: 241) en el contexto de su denodado litigio contra la concepción internista del significado (*infra*), el término "jade" designa de hecho dos minerales distintos. Ambos son inosilicatos, pero si uno es un piroxeno alumínico-sódico, escaso y preciado –la jadeíta–, el otro resulta ser un anfíbole cálcico-magnésico o cálcico-férrico, mucho más común, al que solían atribuirse propiedades medicinales sobre los cólicos del riñón –la nefrita, *lapis nephriticus* o "piedra de ijada", término castellano que en el s. XVI se vertió al francés como "jade". El concepto de jade es, pues, netamente disyuntivo, y consigna propiedades no proyectables: el hecho de que una muestra de jade posea una determinada propiedad, en virtud de la cual podamos contarla como jade, no justifica la inferencia ni constituye evidencia objetiva de que otras muestras de jade también posean dicha propiedad[196].

Al igual que en el caso de la temperatura, de la hidrosolubilidad o de la acidez y alcalinidad (*supra*), esto se toma como indicio de que algo parecido bien puede suceder si tratamos de delimitar con criterios físicos la extensión de tipos de estados psicológicos: a saber, que topemos con que dichos tipos son inherentemente disyuntivos, y las propiedades en cuestión resulten no ser proyectables. Lo más sensato, así pues, sería desistir de la búsqueda de la "forma normal" de descripción psicológica de organismos, y aglutinar nuestros esfuerzos en el desarrollo de teorías neurofisiológicas que den cuenta de los tipos de estados que forman la disyunción. Acaso convenga subrayar que –como bien adelantara Bickle (2006, *supra*)–, la propia tesis de realizabilidad múltiple se ha convertido en manos de Kim en una prueba de cargo contra la viabilidad de una psicología científica autónoma. A los ojos de Kim, en suma, "[…m]ultiple realizability yields the failure of structure-independent mental kinds to meet the standards of scientific kinds" (Bickle 2006: §2.5).

La respuesta de Fodor (1997) al desafío planteado por Kim iría de la mano de la reafirmación –más de veinte años después– de sus planteamientos acerca de la autonomía de las ciencias especiales (Fodor 1974). Es en el carácter indiciario que el caso del jade muestra respecto de los casos en disputa –los tipos de estados mentales, como el dolor– donde Fodor halla la grieta en la que hender la argumentación de Kim. El indicio –viene a decir Fodor– no es tan vehemente que se acerque siquiera a constituir una prueba, porque Kim amalgama impropiamente la realizabilidad variable con la disyuntividad. Se trata, pues, de un *distingo* en sentido puntualmente escolástico: las propiedades disyuntivas, como la de "ser una muestra de jade" no son proyectables ni nomológicas –no forman parte de leyes científicas–, pero las propiedades cuya realización es variable, como tal vez la de "ser un estado de dolor"

[196] No es, desde luego, el único ejemplo de clase natural con estructura química disyuntiva que la mineralogía nos ofrece: de moscovita –$KAl_2(Si_3Al)O_{10}(F,OH)_2$– un silicato conocido también como mica blanca o cristal de Moscovia, se cuentan más de una decena de variedades, unas, como la fuchsita, enriquecidas en cromo –$K(AlCr)_2(Si_3Al)O_{10}(F,OH)_2$–, otras en silicio, en litio o en bario; unas de matices argénteos o verdosos, otras de tono púrpura o rojizo…

–o "sentir dolor", si queremos predicarla del sujeto y no de su estado– son tanto proyectables como nomológicas –precisamente, las leyes científicas en las que aparecen son leyes de ciencias especiales. La analogía entre el jade y los estados mentales, en suma, es espuria (Fodor 1997: 150).

Así que ni "ser una muestra de jade" es una propiedad cuya realizabilidad sea variable, ni es una propiedad que forme predicados proyectables, ni que sea susceptible de intervenir en leyes científicas; todo lo contrario que "ser un estado de dolor". Pero, desde luego, Fodor nos adeuda una motivación de su *distingo* que lo diferencie de un decreto de parte. Alguna razón habrá para que los conceptos mentalistas estén llamados a desempeñar un papel en la psicología más destacado que el que le cabe al concepto de jade en la mineralogía o la gemología, y compete a Fodor señalarla. En otras palabras –las de ellos mismos–, a Fodor le es exigible una respuesta a la pregunta retórica que Kim deja en el aire:

> Why isn't pain's relationship to its realization bases, N_h, N_r, N_m analogous to jade's relation to jadeite and nephrite? If jade turns out to be nonnomic on account of its dual "realizations" in distinct microstructures, why doesn't the same fate befall pain? […] If pain is nomically equivalent to *N*, the property claimed to be wildly disjunctive and obviously nonnomic, *why isn't pain itself equally heterogeneous and nonnomic as a kind?* (Kim 1992: 15)

En este punto, el diálogo entre reduccionistas y antirreduccionistas parece haberse convertido en la reiteración un tanto fatigosa de una recíproca petición de principios, que el propio Fodor describe como una especie de empate dialéctico:

> This is […] a sort of polemical standoff. The functionalist assumes that there are laws about pains "as such", so he infers that, though pain is multiply based, it is *not* (merely) disjunctive. So he infers that pain is unlike jade in the respects that are relevant to the question of projectability. Kim, going the other way around, assumes that pain is (merely) disjunctive, hence that it is relevantly similar to jade, and hence that there are no laws about pain. (Fodor 1997: 153)

El primer ensayo de justificación de la diferente suerte de ambos conceptos que Fodor acomete llega de la mano de consideraciones modales, expresadas en la controvertida jerga lógica de los mundos posibles y la posibilidad metafísica. Así:

> A multiply based property is disjunctive iff it has no realizer in *any* metaphysically possible world that it lacks in the *actual* world. Jade is disjunctive because the only metaphysically possible worlds for jade are the ones which contain either […] [jadeite], or […] [nephrite] or both. By contrast, multiply based properties that are disjunctively *realized* have different bases in different worlds. Pain is disjunctively realized because there's a metaphysically possible, nonactual, world in which there are silicon based pains. (Fodor 1997: 153)

Como Fodor recalca implacable, Kim incurría en *petitio* si diera por asumido que la propiedad de "ser un estado de dolor" no es más que la de ser una u otra de las

diversas encarnaciones del dolor en el mundo actual: eso contraviene la crucial premisa funcionalista según la cual que algo sea un estado de dolor depende de la trama de relaciones causales en que se enrede. Pero no es menos cierto que el propio Fodor *estipula* que ninguna materia que no fuera jadeíta o nefrita, por similar al jade que pudiera resultar en cualquier aspecto, podría ser jade, y obtiene de esa cláusula un rendimiento decisivo para su argumentación: la neutralización de la analogía entre el dolor y el jade.

Bien es verdad que Fodor se esmera en convencernos de que incluso si lográsemos producir en un laboratorio una gema tan similar en todo punto de comparación a la jadeíta y la nefrita como éstas lo son entre sí, difícilmente pasaríamos de concederle el rango de jade artificial. En cambio, si el funcionalista está en lo cierto, generar en un sistema artificial –en un autómata– un estado indistinguible del dolor en los aspectos relevantes –a saber: el patrón de relaciones funcionales con estímulos, respuestas y otros estados internos– sería tanto como haber generado dolor: "[…n]ot *artificial* pains, not pain *simulations*, not *virtual* pains, but *the real things*" (Fodor 1997: 154). La disparidad, sin embargo, descansa sobre suelo menos firme de lo que Fodor necesita: de cara a que algo sea aceptado como una variedad de jade, y por tanto como piedra preciosa, resulta decisiva su *preciosidad*, y eso se compadece mal con la síntesis artificial. De hecho, la relativa abundancia de nefrita lleva a algunos, contra la etimología, a no considerarla *verdadero* jade; otros minerales, como la serpentina o la onfacita, se han empleado como sucedáneos del jade, y no es inimaginable que una eventual combinación de escasez e ignorancia química pudiera haberlos convertido en miembros de la clase delimitada por el concepto de "jade" en algún contexto cultural. Que el jade sea jadeíta-o-nefrita, y nada más, parece más una cuestión de índole pragmática, sujeta a las insondables veleidades de los hechos históricos, que una cuestión de necesidad metafísica o de esencias kripkeanas –más asunto de la esencia nominal que de la presunta esencia real del jade, si queremos decirlo a la manera de Locke. A pie de página, Fodor despacha estas cuestiones taxativamente, reformulando su tesis como la de que "[…] si el jade es jadeíta o nefrita, entonces lo es necesariamente", y, en consecuencia, manifestando una rotunda negativa a "[…] especular sobre lo que haríamos (/deberíamos hacer) si, por ejemplo, encontráramos que hay un tercer tipo de material entre nuestras muestras de jade" (Fodor 1997: 162) –o sea, siguiendo punto por punto el protocolo fijado por Kripke (1972/1980). Pero no acaba de entenderse por qué es legítimo especular sobre el escenario, favorable a la posición de Fodor, en el que alguien genera un material macroscópicamente indistinguible del jade pero que es en realidad, *ex hypothesi*, "cristal de botella fundido" y no sobre el escenario, menos favorable para Fodor, en el que alguien descubre que algunas muestras de lo que venimos considerando jade son en realidad –digamos, también *ex hypothesi*–, un inosilicato molecularmente diferente de la jadeíta y la nefrita, pero suficientemente similar a ambas, y, si se quiere, suficientemente escaso, como para que consintamos (/debamos consentir) en una ampliación de nuestro concepto de jade. Por otro lado, es de rigor reconocer que en la decisión de considerar dolor al

dolor del autómata –es decir, de admitir que *es* dolor–, en cambio, son fundamentalmente consideraciones éticas, a todas luces inconmensurables con las relativas a la carestía de una piedra, que acaso inclinen nuestro juicio del lado que conviene al funcionalista: *in dubio*. Que haya una naturaleza intrínseca del dolor, desligada de tales consideraciones éticas –desligada acaso de cualquier otra consideración, si es que Kripke (1972/1980) está en lo cierto– parece más plausible, en todo caso, que la existencia de una naturaleza intrínseca, e intrínsecamente disyuntiva, del jade.

Un último intento, por parte de Fodor, de persuadirnos de que el jade (si es jadeíta o nefrita) es necesariamente jadeíta o nefrita es comparar el jade con el agua, dando por asumido que el hecho de que el agua sea H_2O es metafísicamente necesario –es decir, que si el agua es H_2O lo es necesariamente, y esa necesidad es, en algún sentido, de índole metafísica. Sin embargo, no parece descabellado aventurar que hay diferencias entre la plausibilidad intuitiva de una ampliación del concepto de jade para abarcar un tercer o cuarto tipo de jade y la plausibilidad intuitiva de una ampliación del concepto de agua para abarcar un segundo o tercer tipo de agua, molecularmente distinto del agua, pero agua al fin y al cabo: no agua oxigenada (H_2O_2: peróxido de hidrógeno), agua pesada (D_2O o ^2H_2O: óxido de deuterio), ni agua superpesada (T_2O ó ^3H_2O: óxido de tritio), sino –valga decir– agua a secas. Y, desde luego, tampoco parece descabellado aventurar que las diferencias en términos de plausibilidad intuitiva entre la ampliación del concepto de jade y la del concepto de agua no han de deberse a que la identidad entre agua y H_2O sea *más* metafísicamente necesaria que la identidad entre jade y $NaAlSi_2O_6$ ó $Ca_2(Mg,Fe)_5(OH)_2(Si_4O_{11})_2$ –o sea, entre jade y jadeíta o nefrita–, sino más bien a que el concepto de agua está mucho más hondamente arraigado en nuestras prácticas, en todos los ámbitos de la vida, que el concepto de jade, lo que lo hace mucho más refractario a cualquier mudanza. Pero decir esto es tanto como recordar, en la estela de Duhem (1906), Quine (1953) o Lakatos (1978), que la distinción entre verdades necesarias y contingentes –como la que en la esfera epistemológica suele irle aparejada, entre enunciados analíticos y sintéticos–, no parece tanto el terso reflejo del hecho de que algunos de nuestros conceptos aprehendan esencias metafísicas irrenunciables, como, más bien, el sublimado de nuestra propensión a proteger de toda vicisitud ciertos conceptos medulares, cuya reforma no podría sino correr pareja a una crisis poco menos que total de nuestra visión del mundo[197].

[197] Entre ellos ocupan un lugar primordial, si Fodor está en lo cierto, las nociones básicas de la psicología de creencias y deseos, impregnadas de intencionalidad y causación mental, que el eliminacionismo amenaza:

> If it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying… if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world. (Fodor 1989: 77; Fodor 1990: 156)

Así que la comparación entre el jade y el agua tampoco es muy fructífera para los intereses de Fodor, si estos pasan por establecer que el jade es metafísicamente idéntico a la disyunción jadeíta-o-nefrita. Pero incluso si acabáramos renunciando a las esencias metafísicas celosamente veladas por Kripke, habremos de convenir en que el papel del concepto de dolor en nuestras prácticas explicativas es sin duda, en cualquier ámbito en que la cuestión se plantee, más afín al del concepto de agua que al del concepto de jade. Es decir: el de dolor se imbrica en la urdimbre de nuestros conceptos tan cerca como el de agua, si no más, de aquello que –refiriéndose, claro está, a un programa de investigación científica– Lakatos (1978) llamaría su núcleo firme, mientras que el concepto de jade, salvo acaso de cara a intereses explicativos muy particulares, ocupa un lugar mucho más periférico que formaría parte, bajo la analogía con el intento de Lakatos de armonizar la visión popperiana de la justificación de las teorías científicas con la descripción kuhniana de su despliegue histórico, del cinturón protector de hipótesis auxiliares.

Desde luego, si las convicciones funcionalistas fuesen verídicas respecto del dolor[198], su naturaleza intrínseca no será intrínseca en todos los sentidos de la expresión, a saber: no será una propiedad básica, de primer orden, de ciertos estados internos o de los sujetos que los padecen, sino una propiedad de segundo orden, la de poseer una propiedad de primer orden cuya caracterización funcional es tarea de la psicología científica confeccionar. En eso, al menos, coinciden la posición antirreduccionista de Fodor y el reduccionismo localizado que Kim aspira a establecer, como el propio Kim se encarga de resaltar:

> The local reductionist must grant that on his view there is nothing intrinsic that all pains have in common in virtue of which they are pains […]. But that is also precisely the consequence of the functionalist view. That, one might say, is the whole point of functionalism: the functionalist, especially one who believes in M[ultiple] R[ealizability], would not, and should not, look for something common to all pains over and above [a certain functional specification] *H* (the heart of functionalism, one might say, is the belief that mental states have no "intrinsic essence"). (Kim 1992: 23)

La apariencia de buena vecindad, sin embargo, es engañosa. Tanto el funcionalista como Kim concuerdan –es cierto– en que los estados mentales carecen de esencias intrínsecas *en ese sentido*: ambos rechazarían la tesis netamente dualista –o, como Kim prefiere plantearlo, emergentista– que se despliega en Kripke (1972/1980) o en Jackson (1982, 1986), y que Kim describe como "la concepción fenomenológica del dolor"[199]. O, dicho de otro modo, los argumentos de Kripke y Jackson resultan, como

[198] En realidad, dicho sea de paso, tanto Fodor (1997) como Block (1997) rechazan el análisis funcionalista del dolor, por razones que ya ambos pusieron sobre la mesa en Block y Fodor (1972, *infra*). Toda la refriega se libra, pues, sobre un bastión abandonado.

[199] Pese a que Kim (2002: 643, *infra*) se muestra complaciente ante la idea de que la faceta fenomenológica de lo mental acabe por eludir tanto el análisis funcionalista como la reducción psicofísica, y permanezca como un reducto epifenomenológico, y pese a que cabría argumentar que en un escenario como ése Kim tendría dificultades para desligarse de una explicación en términos de

es sabido, tan amenazantes para el funcionalismo como puedan serlo para el fisicalismo:

> According to the phenomenological conception of pain, all pains do have something in common: they all *hurt*. But as I take it, those who hold this view of pain would reject any reductionist program, independently of the issues presently on hand: even if there were a species-invariant uniform bridge law correlating pains with a single physical substrate across all species and structures, they would claim that the correlation holds as a brute, unexplainable matter of fact, and that pain as a qualitative event, a "raw feel", would remain irreducibly distinct from its neural substrate. (Kim 1992: 22)

En cambio, en un sentido diferente e igualmente legítimo de un atributo tan deslustrado como es "intrínseco", el funcionalista concede al dolor una naturaleza intrínseca que Kim le niega –una esencia, hasta donde el funcionalista puede conceder que los estados mentales tengan tales esencias[200]–: precisamente su naturaleza funcional. La insistencia de Fodor en que existen mundos posibles en los que hay estados de dolor cuya encarnación física es por completo diferente de la de los nuestros es simétrica de su convencimiento de que no existen mundos posibles en los que haya estados de dolor cuya caracterización funcional sea diferente de la de los nuestros. Tal es precisamente –conviene recordarlo– la certidumbre que Lewis (1980) trataría de socavar con su conocido ejercicio modal en torno al *dolor marciano*, de diversa encarnación física, y el *dolor insensato*, de diversa caracterización funcional –un dolor causado por "[…] el ejercicio físico moderado con el estómago vacío" y que ocasiona facilidad para concentrarse en tareas matemáticas, así como cierta tendencia a cruzar las piernas y chasquear los dedos (Lewis 1980: 216, *cf.* también Gunderson 1971, *infra*). Pero el mero planteamiento de la posibilidad de concebir el dolor insensato sería una ociosa contorsión argumental de no quedar trazado ante un adversario dialéctico que otorgara rango de necesidad a la identidad entre el dolor y un cierto patrón funcional, haciendo así de dicho patrón, en algún sentido, la esencia del dolor.

Así que la apelación por parte de Kim a la afinidad que a este respecto guardan el funcionalismo y su reduccionismo localizado no es sino una poco disimulada *ignoratio elenchi*. Dicha afinidad, aunque existe, no es relevante en la controversia: el propio Kim lo hace ver implícitamente cuando señala que el hecho de que quienes apuestan por una concepción fenomenológica de lo mental rechacen el acuerdo entre funcionalistas y reduccionistas es independiente de las cuestiones que articulan dicha controversia. Cuando se mira de frente a esas cuestiones, la respuesta, o más bien las respuestas, de Kim (1992) a la pregunta que Block (1980: 178-179) le habría acusado de esquivar –"What is common to the pains of dogs and people (and

---

propiedades fenomenológicas intrínsecas, aunque epifenoménicas, del hecho, difícilmente controvertible, de que existan *clases* de sensaciones: si tales clases no se forman ni por criterios funcionales ni por criterios fisiológicos o físicos, el criterio fenomenológico se perfila como último y acaso inevitable recurso.

[200] *Cf.* Braddon-Mitchell y Jackson (1996: 101).

all other species) in virtue of which they are pains?"– resultan tan poco satisfactorias como la de Lewis (1969: 233, *infra*): una "estrategia desesperada", como el propio Lewis (1980: 217, *infra*) diría acerca de otra respuesta a la misma cuestión.

Pero antes de adentrarnos otra vez en esos parajes, es preciso retomar el hilo de la argumentación de Fodor sobre las diferencias entre el dolor y el jade. No parece que podamos darnos por convencidos de que la analogía que entre ambos plantea Kim no es "sólida e instructiva", como piensa el propio Kim (1992: 17), sino "espuria", como sentencia Fodor (1997: 150). La tesis de Fodor de que el jade es necesariamente la disyunción jadeíta o nefrita, pero el dolor no es necesariamente, aunque pueda serlo de hecho, la disyunción N_h o N_r o N_m es endeble, porque se estriba –decíamos– sobre cimientos desparejos. El carácter contingente de la identidad entre dolor y la disyunción de sus encarnaciones físicas es suelo razonablemente firme, sobre el que conviven la concepción funcionalista de lo mental y la concepción fenomenológica: es en Kim, en todo caso, en quien recaería el peso de la prueba si hubiera de afirmar que esa identidad es necesaria. El carácter necesario de la identidad entre el jade y la disyunción de jadeíta y nefrita, en cambio, es terreno inestable, y es Fodor quien debe atestiguar en su defensa. Si hemos de esquivar la conclusión de Kim, en fin, se impone tantear otras rutas.

Reconstruir la diferencia entre la identidad de agua y H_2O, por un lado, y la jade y $NaAlSi_2O_6$ ó $Ca_2(Mg,Fe)_5(OH)_2(Si_4O_{11})_2$, por otro, bajo el prisma lakatosiano de la distancia entre el núcleo firme y el cinturón protector de un programa de investigación científica –o, en este caso, más bien, de una cierta visión del mundo derivada del programa de investigación de la química analítica– deja en el aire varias cuestiones cruciales. Está por ver, en primer lugar, si la identidad entre el dolor y el patrón funcional característico del dolor arraigaría, desde esa perspectiva, más cerca del núcleo de nuestra concepción de lo mental –ya de nuestra concepción espontánea de lo mental, ya de la derivada del programa de investigación de las diversas ciencias de la mente y el cerebro, si es que cabe una distinción nítida entre ambas– que la identidad entre dolor y la disyunción de estados cerebrales que puedan instanciar dolores particulares en distintos sujetos e instantes; dicho de otro modo, si el dolor insensato nos resulta inconcebible en algún sentido en el que no lo sea también el dolor marciano. Ahora bien, no resulta en modo alguno aventurado dar por bueno –como acaba de hacerse– el dictamen según el cual la identidad entre dolor y la miscelánea fisiológica que lo concita es de naturaleza contingente incluso para los más enérgicos defensores del reduccionismo psicofísico. Por otra parte, que el vínculo entre el dolor y ciertos estímulos, conductas y estados internos resulte o no ser necesario forma parte del corazón de la controversia: es –también acaba de quedar recalcado– la cuestión sobre la que trata de incidir Lewis (1980), quien se apresura a adelantar su convicción de que "[…] el dolor sólo se halla asociado contingentemente con su rol causal" (Lewis 1980: 217). Sin embargo, a efectos de ubicar la relación entre el dolor y su propia especificación funcional en el espacio metateórico que separa al núcleo firme y al cinturón protector de un programa de investigación científica –o algo parecido– basta con menos que con un certero

refrendo de su carácter necesario. No en vano, es precisamente la extensión de dicho espacio como un continuo, más dúctil que los dogmas empiristas denunciados por Quine (1959), lo que nos ha inclinado a reconstruir en términos lakatosianos la distinción discreta entre verdad necesaria y contingente, o al menos un segmento de ella. Sólo es menester, entonces, para aproximar la identidad entre el dolor y su especificación funcional al núcleo firme de nuestra concepción del dolor, constatar que el supuesto carácter necesario de la identidad entre el dolor y su propia especificación en términos funcionales resulte menos controvertido que el supuesto carácter necesario de la identidad entre el dolor y su propia especificación en términos neurofisiológicos. Pero eso es, efectivamente, lo que se acaba de constatar: que ambas partes asumen que la identidad entre dolor y determinado estado cerebral es de índole contingente, mientras que la naturaleza necesaria o contingente de la identidad entre dolor y determinado patrón de relaciones funcionales es medular a la controversia. Así, pongamos por caso, al convencimiento de Lewis (1980) de que puede existir un dolor insensato, imbricado en una urdimbre causal totalmente ajena a la de nuestro dolor, es fácil oponer la persuasión expresada por Putnam (1967a: 229, *supra*) respecto a que simplemente rechazaríamos que el concepto de dolor se aplique a un organismo cuyo estado no venga causado por estímulos semejantes a los que en nosotros suelen causar dolor, ni cause conductas de evitación, etc., del mismo modo que nos resistiríamos a aceptar que un animal esté sediento si su conducta no parece orientada a la saciación mediante la ingesta de líquido[201]. Incluso en el terreno de una lectura analítica del funcionalismo afín a la tesis de identidad psicofísica, Braddon-Mitchell y Jackson (1996) admiten de buen grado, en el contexto de su defensa de los méritos del conductismo lógico, que:

> In sum, it is of the essence of mental states to show up in behavior in appropiate circumstances: a person in pain is disposed to move in such a way that the pain is relieved; intelligent people are better at solving problems than unintelligent people […]. What is more, acknowledging these connections is part of understanding what it is to be in a mental state. Someone who does not realize that if Helena is better at chess than Mario, then Helena typically beats Mario when they play chess against each other, does not properly understand the concept of being good at chess. (Braddon-Mitchell y Jackson 1996: 31)

Aunque Braddon-Mitchell y Jackson eluden mencionar las reflexiones de Lewis (1980) –a quien de hecho reverencian como "[…] uno de los principales arquitectos del funcionalismo de sentido común" (Braddon-Mitchell y Jackson 1996: 54)–, parece difícil que pudieran eludir también la conclusión de que el dolor insensato a duras penas sería reconocible como dolor, o de que no entiende cabalmente el concepto de

[201] *Cf*. Wittgenstein (1953: §281):

> "¿Pero lo que tú dices no viene a ser que no hay, por ejemplo, ningún dolor sin conducta de dolor?" –Viene a ser esto: sólo de seres humanos vivos y de lo que se les asemeja (se comporta de modo semejante) podemos decir que tienen sensaciones, ven, están ciegos, oyen, están sordos, son conscientes o inconscientes.

dolor quien lo asemeja al dolor insensato. En realidad, no podía ser de otro modo toda vez que, desde la óptica del funcionalismo analítico la descripción del patrón de relaciones con estímulos, respuestas y otros estados mentales típico de un estado mental –"típico", porque la descripción se acompaña de cláusulas *cæteris paribus*– proporciona el *significado* del término con que designamos al estado mental en cuestión (*cf*. *infra*).

Como ya se ha comentado, pertenece a la propia posición dialéctica de los argumentos de Lewis (1980) el reconocimiento de cierto consenso en cuanto al carácter necesario del lazo entre estados mentales y sus caracterizaciones funcionales –al menos, en particular, en el caso del dolor–, en la medida en que resquebrajar dicho consenso funcionalista es parte del propósito del recurso a la posibilidad del dolor insensato. Resulta significativo, además, que Lewis desista de fundamentar dicho recurso refugiándose en un vaga declaración de desconocimiento de los medios para probar una posibilidad: "[…n]o sé cómo probar que algo es posible", dice Lewis (1980: 216). De hecho, el camino por el que transita Lewis para inculcarnos que el dolor insensato es posible es tan conocido como, aunque Hume lo presentara como "[…] an establish'd maxim in Metaphysics"[202], reñida su fiabilidad: se trata de convencernos de que el dolor insensato resulta concebible, y bascular sobre la discutible premisa implícita de que todo lo concebible es posible. Ahora bien, la fuerza de la apelación de Lewis (1980: 216) a que el dolor insensato resulta concebible descansa a su vez, fraudulentamente, sobre una concepción fenomenológica del dolor que el propio Lewis, al igual que Kim (1992: 22, 23, *supra*), impugna sin contemplaciones, pero que en ese contexto elude desautorizar, o mencionar siquiera. Dicho de otro modo: asentimos, pensamos que el dolor insensato es posible, porque asumimos que al insensato, cuando hace ejercicio con el estómago vacío y siente un irrefrenable deseo de cruzar las piernas o concentrarse en la realización de ejercicios matemáticos, *le duele*, donde el giro "le duele" se refiere al *quale*, al carácter cualitativo o vivencial, a la fenomenología de un estado interno del sujeto en cuestión, a una propiedad intrínseca, no relacional de dicho estado interno –a saber: que *es doloroso*. Pero tales atributos cualititativos intrínsecos son tan disonantes en la concepción funcionalista del dolor, o de la mente en general, como en la concepción fisicalista, y buscar un contrafuerte en ellos que sirva para dirimir la polémica entre las interpretaciones reduccionista y antirreduccionista del funcionalismo sería, en consecuencia, flagrantemente ilegítimo.

Hay razones firmes, así pues, para concluir que la identidad entre el dolor y su patrón funcional característico habita más cerca del núcleo de nuestra concepción de lo mental que la identidad entre dolor y la disyunción de estados cerebrales que puedan instanciar dolores particulares en distintos sujetos e instantes. Pero esto –ya sabemos– es en todo caso una conclusión sobre el concepto de dolor, que no es seguro que ataña al dolor en tanto que propiedad. Así que el mismo razonamiento

[202] Sigue: "[…] that whatever the mind clearly conceives includes the idea of possible existence, or in other words, that nothing we imagine is absolutely impossible" (Hume 1739: §1.2.2.8)

que Putnam (1967a, *supra*) empleara para bloquear una deriva hacia el conductismo de sus reflexiones sobre el modo en que identificamos nuestros estados mentales bloquea ahora la extracción de conclusiones antifisicalistas de la distancia que, en el seno de nuestra comprensión más o menos intuitiva de lo mental, pueda separar al patrón funcional típico del dolor y a sus encarnaciones fisiológicas. En suma, esta vía no conduce mucho más allá del empate dialéctico ya trazado por Fodor (1997: 153, *supra*), pues queda por ver si da razón de alguna manera a la lectura antifisicalista del funcionalismo ante su lectura reduccionista.

Dotar a su *distingo* entre propiedades de realización variable –proyectivas, nomológicas– y propiedades disyuntivas –que no son ni una cosa ni la otra– de puntales lógicos en los que Kim se vea obligado a convenir, y así esquivar esa podredumbre del diálogo, es sin duda lo que pretende Fodor (1997) al introducir la diferenciación entre disyunciones abiertas y disyunciones cerradas –que tanto Block (1997) como Jacob (2002), por lo demás reacios a las conclusiones de Kim, estimarían inservible. De lo que se trata es, entonces, de elucidar *por qué* las propiedades disyuntivas, como es según Fodor la de ser una muestra de jade, resultan no ser proyectables, y en cambio las propiedades de realización variable, como la de ser una instancia de dolor, sí lo son. La tarea se vislumbra crucial porque –recordemos:

> Functionalists are required to deny that pain is *identical to* the disjunction of its realizers. The reason they are is that it's part of their story that the functional property realized, *but not its physical realizer*, is projectible. And the reason they have to say *that* is that *otherwise multiple realization wouldn't be an argument against reduction* […] (Fodor 1997: 155)

Aparte de crucial, la tarea es más delicada de lo que acaso parezca a primera vista. El funcionalista asegura que existen leyes psicológicas sobre el dolor irreductibles a leyes neurofisiológicas sobre la disyunción de estados nerviosos que lo encarnan. Si se pliega a la estrategia de Fodor, ha de negar también tanto que existan leyes sobre el jade como que las haya acerca de la disyunción jadeíta-o-nefrita. Lo que debe justificar es la asimetría de su posición: si una disyunción que Fodor considera metafísicamente cerrada, como jadeíta-o-nefrita, no es proyectable ni nomológica –no es capaz de sustentar ley alguna–, y tampoco lo es una disyunción que Fodor considera metafísicamente abierta, como la de las posibles encarnaciones del dolor, ¿por qué en el primer caso se niega el carácter proyectable y nomológico de la propiedad disyuntivamente instanciada –ser una muestra de jade– y en cambio se otorga carácter proyectable y nomológico a la propiedad instanciada en el segundo caso de forma disyuntiva o, si se quiere, múltiple –ser una instancia de dolor? A estas alturas, Fodor parece forzado a afirmar que la razón de que las disyunciones cerradas no sean proyectables no es la misma que la de que no lo sean las disyunciones abiertas: acaso eso le permitiera rendir cuentas del diferente estatus que concede a las propiedades instanciadas en disyunciones cerradas y en disyunciones abiertas en cuanto a su carácter proyectivo y nomológico. Ahora bien, la razón por la que las disyunciones cerradas no son proyectables es sencillamente que la

constatación en uno de sus elementos de la propiedad que lo convierte en miembro de la disyunción no justifica –o no permite objetivamente, *cf.* Block (1997, *supra*)– la atribución de esa propiedad a otro elemento de la disyunción. Si la razón por la que las disyuciones abiertas no son proyectables no es precisamente ésa, ¿cuál es?

La respuesta de Fodor (1974) intentaba ser expeditiva: una disyunción abierta podría resultar proyectable *prima facie* en la medida en que tuviéramos en cuenta su mera aparición en las leyes-puente que la enlazan con la propiedad que se instancia en ella; no obstante, debemos exigir que esa presencia en las leyes-puente venga avalada por leyes formuladas en el mismo nivel explicativo en que se formula la propia disyunción y corroboradas independientemente. La razón de que debamos exigir tal cosa es que sólo así podremos eludir la acumulación de leyes-puente que incurran en la arbitraria heterogeneidad del *gerrymandering*[203], en las que "[…] intuitively, *all* that the disjuncts have in common is that they realize some higher level state" (Fodor 1997: 156). Sería, pues, este veto a la práctica del *gerrymandering* en la postulación científica de clases naturales lo que haría improbable que las disyunciones abiertas que instancian estados funcionales alcanzaran una genuina proyectabilidad, o, dicho de otro modo, lo que haría poco prometedor el fisicalismo de tipos dejando intacta la plausibilidad del fisicalismo de casos:

> On this account, the constraints on bridge laws are *stronger than* (in fact, they include) the constraints on proper (single-level) laws. This is what underlies the intuition that type materialism comes to more than just the claim that it's nomologically necessary that every nonbasic property be physically realized. (Fodor 1997: 157)

El terreno compartido con Kim es, entonces, que las clases disyuntivas no son proyectables en ningún caso, con independencia de lo arbitrario de su formación. Pero a juicio de Kim el carácter no proyectable de las clases disyuntivas tiene causas uniformes: sencillamente, la proyección de las propiedades relevantes de determinados elementos de la clase se ve refutada por otros elementos. Según Fodor (1974), en cambio, la imposibilidad de proyectar las propiedades de las disyunciones abiertas –las que instancian propiedades de orden superior cuya realizabilidad resulta ser variable– es un caso especial: lo que las hace refractarias a la proyección no es tanto que la disyunción quede abierta como que se ha formado de manera arbitraria, es decir, que su única aparición en leyes científicas ocurre precisamente en las leyes-puente que dan cuenta de su relación con la propiedad de orden superior. Al cabo de los años, sin embargo, el carácter abierto de las disyunciones en las que se instancian propiedades de realización variable vuelve a aparecer ante Fodor (1997) como la piedra angular de su ineptitud para sustentar leyes científicas, aunque no

[203] Por Elbridge Gerry, quinto vicepresidente de los Estados Unidos (1813-1814), quien durante su mandato como gobernador de Massachusetts, impulsó una ley que redistribuía los distritos electorales del estado de forma descaradamente beneficiosa para los intereses de su partido. Comoquiera que el mapa electoral resultante remedara una salamandra, el editor del *Boston Gazette* lo bautizó como "gerrymander", crasis que hizo fortuna en el discurso político, hasta hoy.

hay una deserción del análisis inicial. El problema con las leyes que contienen disyunciones abiertas –dirá Fodor (1997)– es simple y llanamente que "[…] sugieren generalizaciones perdidas":

> To offer a law of the form $R_1 \vee R_2 \vee \ldots \rightarrow Q$ is to invite the charge that one has failed correctly to identify the property in virtue of which the antecedent necessitates the consequent. […] Someone who offers such a law undertakes a burden to provide positive reason that there isn't a higher level but nondisjunctive property of things that are $R_1 \vee R_2 \vee \ldots$ in virtue of which they bring it about that Q. (Fodor 1997: 158)

Es más: nuestro anhelo de llegar a formular cada ley en la que pudiera aparecer una disyunción abierta en términos de una propiedad de nivel superior que abarque los términos de la disyunción *deriva su legitimidad de la naturaleza misma del razonamiento inductivo*. Se trata, en definitiva, del mismo anhelo que nos impulsa a preferir una generalización universal a una enumeración abierta de casos, aún a costa de hipostatizar la propiedad que soporta la generalización. Bien podrían existir leyes sobre disyunciones abiertas que no pudieran ser abarcadas mediante propiedad de orden superior alguna –concluye Fodor–, pero ello no es óbice para que tengamos "[…] general methodological grounds for preferring a closed law to a corresponding open one, all else equal", puesto que "[…] this policy complies with an injunction that all of our inductive practice illustrates: *Prefer the stronger claim compatible with evidence, all else equal*" (Fodor 1997: 159). Una moderada reificación de las propiedades funcionales, fundada al fin y al cabo en la estructura inductiva del conocimiento científico, se perfila desde esta perspectiva como el sustrato del antirreduccionismo y el realismo acerca de lo mental del que Fodor, y con él buena parte de la ortodoxia cognitivista, han venido haciendo gala desde que se levantaran las primeras voces contra el conductismo. Lo que Lewis (1969: 233, *infra*) impugnara como el anticuado prejuicio según el cual los nombres de entes necesarios deben designarlos necesaria y absolutamente vendría a ser, bajo una mirada más benigna, nada más que esta propensión a admitir la hipóstasis de aquellas propiedades que acrediten su cometido en enunciados legaliformes verdaderos –algo que, al hilo de la conocida afirmación de Strawson (1952) sobre los principios y prácticas inductivas que empleamos, bien cabría considerar "[…] constitutivo de nuestro concepto de racionalidad".

Ha quedado ya anticipado que ni Block (1997) ni Jabob (2002) estiman concluyentes las objeciones de Fodor (1997) al argumento de Kim (1992) que es el germen de la controversia. En efecto, Block considera que la distinción relevante no es la invocada por Fodor con la ya sabida letanía: propiedades disyuntivas, cerradas, *versus* propiedades de instanciación variable, abiertas. A su entender, la posición de Fodor se condensa en la propuesta de añadir al canon nageliano de reducción interteórica el requisito de que las leyes-puente deben enlazar entre sí *clases naturales*, más la asunción de que las disyunciones abiertas no son clases naturales. Falta –dice Block (1997: 112)– un razonamiento que nos persuada de que "[…] bridge principles that link kinds to heterogenous disjunctions are importantly defective". Sin embargo,

la objeción de Block al argumento que Fodor ofrece en ese sentido no es muy vigorosa. El problema de las disyunciones abiertas sobre las que gira la controversia, según Fodor, es que su condición de clases no está certificada independientemente de su aparición en la ley-puente en cuestión, pues no forman parte de leyes en el vocabulario de la disciplina reductora. Esa condición –replica Block– es inocua: si $P_1 \rightarrow P_2$ es una ley psicológica, y si tanto $P_1 \rightarrow (N_1 \vee N_1^*)$ como $P_2 \rightarrow (N_2 \vee N_2^*)$ son leyes-puente que la enlazan con la ley neurológica $(N_1 \vee N_1^*) \rightarrow (N_2 \vee N_2^*)$, cuyos términos se refieren ambos a disyunciones abiertas[204], debemos dar la razón a Kim en que $(N_1 \vee N_1^*) \rightarrow (N_2 \vee N_2^*)$ ostenta en el mismo grado la propiedad de necesidad nómica que $P_1 \rightarrow P_2$. Si la neurología no contenía la ley $(N_1 \vee N_1^*) \rightarrow (N_2 \vee N_2^*)$ antes del proceso de reducción interteórica, sólo hay que agregársela, y la reducción está cumplida *pace* Fodor. Además, Fodor –piensa Block– no puede discutir que $(N_1 \vee N_1^*) \rightarrow (N_2 \vee N_2^*)$ sea una ley, porque eso es tanto como discutir que $(N_1 \vee N_1^*)$ y $(N_2 \vee N_2^*)$ sean clases naturales, con lo que incurriría en petición de principios. Es Block, sin embargo, quien desliza una falacia en el argumento: *ignoratio elenchi*. Lo que Fodor exige no es que $(N_1 \vee N_1^*)$ y $(N_2 \vee N_2^*)$ aparezcan en *alguna* ley de N –una cualquiera–, sino que aparezcan en una ley de N que sirva de certificación *independiente* de su condición de clase natural: es decir, *que no exprese la misma regularidad que* $P_1 \rightarrow P_2$. De hecho, $(N_1 \vee N_1^*) \rightarrow (N_2 \vee N_2^*)$ es la única ley de N que ni cumple esa condición ni puede desempeñar esa función –excepción hecha, por supuesto, de inagotables caprichosas construcciones lógicas coextensivas del estilo de $[(N_1 \vee N_1^*) \wedge N_1 = N_1] \rightarrow [(N_2 \vee N_2^*) \wedge N_2 = N_2]$. Así que en esto, según parece, la razón asiste a Fodor. Acierta Block, en cambio, cuando insiste en que, aun si esto se concede a Fodor, seguimos adoleciendo de una respuesta a la pregunta de por qué el hecho de que las disyunciones abiertas no sean clases naturales las inhabilita para formar parte de leyes-puente. Pero el propio Block había esbozado la respuesta cuando, unas líneas antes, menciona otro requisito inherente al concepto de reducción: se espera que las leyes de la teoría reductora, unidas a las leyes-puente, *expliquen* las leyes de la teoría reducida. Como él mismo apunta con lucidez:

> This condition is often ignored in the debate over multiple realizability because of the widespread positivist assumption that explanation is just deduction. If the terms of the upper level theory are all definable in lower level terms, explanation of the upper level laws is said to be trivial. The upper level laws can be deduced from the lower level theory plus definitions, and if the lower level theory isn't rich enough, the "images" of the upper level laws can simply be added to the lower level theory. (Block 1997: 111)

La denuncia de ese prejuicio era clave en el ejemplo del cilindro y la tabla (Putnam 1975, *supra*), que Block cita ignorando precisamente ese aspecto –también, por cierto, resulta decisiva en Pylyshyn (1984). Bajo este prisma, la lectura que del planteamiento de Fodor (1997) se nos impone es, en síntesis, que las disyunciones

[204] Aunque Block supone, por mor de la simplicidad, que se trata de disyunciones con sólo dos elementos.

abiertas no pueden formar parte de leyes-puente porque no son clases naturales, y las clases que no son clases naturales no pueden formar parte de leyes-puente porque el resultado fracasaría a la hora de *explicar* las leyes de la teoría reducida, aunque permitiera deducirlas. Si Block no encuentra convincente el razonamiento de Fodor en cuanto a las razones de ese fracaso –es decir, si no ve obvio en qué sentido la exigencia de adecuación explicativa podría persuadirnos de que "[…] bridge principles that link kinds to heterogenous disjunctions are importantly defective"–, es porque, a lo que parece, ignora el requisito de independencia que Fodor impone sobre las leyes de la teoría reductora que hayan de certificar la condición de clase natural de la disyunción sobre la que se reduce una clase natural descrita en el vocabulario de la teoría reducida. Además –añadiría Fodor– una disyunción abierta, al dejar en el aire una generalización fallida, violenta el armazón mismo de nuestra práctica inductiva.

De todos modos, aun si se concede a Fodor que la heterogeneidad de las disyunciones abiertas las hace indóciles en lo que concierne a la proyección de propiedades, Block estima que su esforzado embate deja indemnes las conclusiones de Kim, o incluso las refuerza. La razón –ya sabemos– es que tan indóciles como sean tales disyunciones habrán de ser también los tipos de estados psicológicos que les son nomológicamente coextensivos: "[…] pain is just as non-projectible and non-kind-like as that open disjunction" (Block 1997: 116). Pero parece haber caído en el olvido, en este punto, que ésa es precisamente la tesis de Kim que Fodor discute, insistiendo –recordemos– en que "[…] la propiedad funcional, *pero no su realización física*, es proyectable" (Fodor 1997: 155, *supra*): invocarla, así pues, no pasa de ser una franca petición de principios. Lo que subyace a la falacia, por lo demás, parece ser el presupuesto de que cualquier generalización puede expresarse adecuadamente en el vocabulario teórico designado como propio de las ciencias básicas, presupuesto que no es arriesgado describir precisamente como una secuela de la confusión positivista entre explicación y deducción que Block censura.

Otra faceta del mismo presupuesto, por cierto, se hace patente cuando Block (1997: 116) infiere que las disyunciones heterogéneas no son clases naturales a partir del hecho de que sus elementos no están ligados por una relación de *similitud*, que sería el fundamento de la noción de clase natural[205]. El problema, obviamente, es que no hay razones sólidas para suponer que si la relación de similitud se da entre los elementos de una clase bajo una descripción designada como básica, se dará también bajo cualquier otra descripción, y que no se dará bajo ninguna descripción si no se da en la descripción básica: o, más rotundamente, acaso haya razones sólidas para suponer que la relación de similitud es relativa al vocabulario en que se formule la

[205] El argumento de Block es en realidad algo más confuso: a su entender, que una propiedad no sea proyectable *causa* que no sea una clase natural; asimismo, el *fundamento* de las clases naturales es la similitud, de la que carecen –por definición, cabe añadir– las disyunciones heterogéneas. Lo que no queda claro es qué relación exista entre su *carácter no proyectable* como causa de que una propiedad no constituya una clase natural y la *falta de similitud* entre sus elementos como *fundamento* de esa misma condición. Tal vez se trate, a ojos de Block, de una distinción sin diferencia.

descripción. Por aprovechar el ejemplo aducido por el propio Block, el bromuro potásico (KBr) y el hidrato de cloral ($CCl_3CH(OH)_2$) no albergan ninguna similitud entre sí bajo el prisma de la formulación química, pero sí bajo el de la psicofisiología: ambos son sedantes –"Cuius est natura / Sensus assoupire" (Molière 1673: III, *supra*). Pero esto es una de las tesis capitales de la vertiente epistemológica del funcionalismo: de lo que se trata –nos decía Pylyshyn (1984, *infra*)– es de la capacidad de distintos vocabularios para capturar distintas generalizaciones, y, por tanto, de la idoneidad de las explicaciones formuladas en tales vocabularios. La objeción de Kim, que Block parece secundar, es que la propiedad de tener efectos sedantes no puede ser una clase natural porque no es proyectable: no podemos, por ejemplo, inducir que todas las sustancias sedantes son cancerígenas aunque todas las muestras que hayamos analizado lo sean, porque pudiera darse el caso de que todas nuestras muestras lo fueran de bromuro potásico, siendo así que –supongamos– el hidrato de cloral fuese inocuo desde el punto de vista de la carcinogénesis. Otra posible interpretación de la improcedencia de tal paso inductivo, sin embargo, sería que el vocabulario teórico de la psicofisiología y el de la oncología son estancos en lo relativo a la posibilidad de proyectar generalizaciones relativas a la clase natural de las sustancias sedantes y a la clase natural de las sustancias cancerígenas. Por supuesto, se trata de una interpretación menos afín al ideal positivista de la ciencia unificada, pero es que uno de los trabajos seminales de la concepción funcionalista de la explicación psicológica, precisamente el de Fodor (1974), no es sino un alegato contra ese ideal. Desde un ángulo ontológico, por lo demás, la oposición de Kim a los principios epistemológicos del funcionalismo se sublima –como se ha visto– en su estricta adherencia a un principio de herencia causal (Kim 1993a: 355, *supra*) derivado de la idea de clausura física: toda eficacia causal corresponde a las propiedades físicas, y toda relación causal en que parezca intervenir una propiedad funcional, cuya encarnación pueda resultar variable, es en realidad heredada de una relación causal en la que interviene la encarnación física de dicha propiedad funcional. Puesto que la explicación atañe a la delimitación de clases naturales, y éstas –entiende Kim– se forman a tenor de los poderes causales de sus elementos, se sigue que las únicas clases naturales que existen son las deslindadas por la física: la física, al hilo de la fértil metáfora de *Fedro* 265e (*supra*), es el único carnicero capaz de desmembrar la naturaleza por sus articulaciones[206].

Con todo, la réplica adecuada a Kim es según Block un *distingo* –ahí atinaría Fodor–, aunque otro: el que cabe trazar entre las propiedades de instanciación de los estados mentales –aquéllas que un estado mental posee en virtud de peculiaridades de su encarnación en un sistema en particular, para las que el razonamiento de Kim sería válido– y propiedades *D* de los estados mentales –aquéllas cuya instanciación está supeditada a ciertas restricciones físicas y evolutivas, y que pueden resultar proyectables, *pace* Kim, aun cuando sus encarnaciones físicas difieran entre sí de

[206] En estos términos, el propio Block (2003) disputará como veremos los planteamientos de Kim con argumentos que confluyen parcialmente con los esgrimidos por Davidson (1993).

modo que hagan la disyunción de todas ellas refractaria a la proyección. La noción de propiedades *D* –donde "*D*" está por "diseño" y, más jovialmente, por "Disney"– parte en el trabajo de Block de una candorosa reflexión sobre el cine de dibujos animados: que "[…] las leyes de la naturaleza imponen restricciones sobre los modos de hacer algo que satisfaga una cierta descripción" se hace patente tan pronto como reparamos en que en las películas de Walt Disney, "[…] las tazas de té piensan y hablan, pero en el mundo real, cualquier cosa que pueda hacer eso precisa más estructura de la que tiene una taza de té. Podríamos bautizar esto como el Principio de Disney […]" (Block 1997: 120). Por cándida que pueda parecer, la admisión de este principio no es trivial en el seno de la tradición funcionalista, en cuyas actas fundacionales –recordemos– consta la osada afirmación de Putnam (1975b: 291, *supra*) según la cual "[…w]e could be made of Swiss cheese and it wouldn't matter". Casi como quien presenta sus disculpas añade Block a continuación que:

> It is easy to be mesmerized by the vast variety of different possible realizations of a simple computational structure, say that of an *and* gate, which can be made of cats, mice and cheese (Block 1995) as well as mechanical or electronic components. But the vast variety might cut down to very few when the function involved is mental, like thinking, for example, and even when there are many realizations, laws of nature can impose impressive constraints. (Block 1997: 120)

El propio Block (1990c: 39), en efecto, había argumentado no mucho antes –y Searle (1992: 206) lo había citado a modo de ejemplo del desprecio funcionalista hacia lo biológico–, con relación a las compuertas lógicas que reproducen los operadores booleanos en los circuitos electrónicos de conmutación, que –tal como lo expresaría poco después:

> How such gates work is no more part of the domain of cognitive science than is the nature of the buildings that hold computer factories. This reveals a sense in which the computer model of the mind is profoundly un-biological. We are beings who have a useful and interesting biological level of description, but the computer model of the mind aims for a level of description of the mind that abstracts away from the biological realizations of cognitive structures. As far as the computer model goes, it does not matter whether our gates are realized in gray matter, switches, or cats and mice. (Block 1995b: 385-386)

Conviene no sobreestimar, con todo, el alcance de esta putativa retractación, y la forma más eficiente de *sí* sobreestimarla es sobreestimar, de partida, el de la tesis de realizabilidad múltiple. El compromiso funcionalista es que sistemas idénticos desde un punto de vista funcional pueden ser indefinidamente diferentes en lo que atañe a su composición material –de qué y cómo estén compuestos– *mientras* ciertas relaciones funcionales se vean preservadas y, por añadidura, que ese punto de vista funcional es el idóneo de cara a la explicación psicológica –sea, como veremos, a la elaboración de teorías psicológicas empíricas o a la reconstrucción del discurso psicológico ordinario. Si se prescinde de la cláusula condicional, la doctrina

funcionalista queda exagerada y se convierte en un espantapájaros. Lo que Block apunta al relatar el efecto casi hipnótico de las primeras ejemplificaciones de la idea de máquina abstracta es, bien entendido, que la cláusula condicional acaso sea de mucho más difícil satisfacción, en el caso de procesos psicológicos complejo, de lo que aquellos primeros pasos de la teoría de la computación pudieron hacernos imaginar. Pero si hubiéramos desquiciado el sentido original de la tesis de realizabilidad múltiple desproveyéndola de la cláusula condicional, las palabras de Block adquirirían casi el aire de una apostasía[207].

De cualquier manera, aun a sabiendas de que no entraña un descabalgamiento de las tesis primordiales del funcionalismo, es obvio que al poner de relieve las limitaciones estructurales que constriñen el rango de posibles implementaciones físicas de un sistema funcionalmente descrito –ya sean éstas de índole física o química, ya resulte preciso describirlas en un vocabulario más abstracto[208]–, Block está al menos dando un golpe de timón a bordo de cierta tradición cognitivista. A esas restricciones estructurales se agregarían, además, en distintos casos y proporciones, varias "fuerzas", tales como la selección natural, los procesos de aprendizaje, o la intervención de un diseñador, cuya tendencia homogeneizante cobraría vigor al verse angostado su caudal a los cauces –los "canales", dice Block– provistos por las propias restricciones estructurales.

La naturaleza dual de los factores que limitan la heterogeneidad de las implementaciones posibles de un sistema funcional sirve entonces a Block para abastecer de nuevas guarniciones el núcleo doctrinal del funcionalismo –del que, como anticipábamos, no hay abdicación. Así, Block no se demora en constatar que, si bien estas consideraciones apuntan a que no deberíamos esperar una heterogeneidad total ni en el nivel del diseño ni en el de la implementación[209] también apuntan a que

[207] Como veremos más adelante, se ha asumido en ocasiones que la autonomía de la explicación psicológica que pueda derivarse de la tesis de realizabilidad múltiple equivale a la irrelevancia de una aproximación neurofisiológica, lo que ha contribuido en poner en guardia contra el funcionalismo a quienes partían de una mayor afinidad con dicha aproximación. Pero una cosa es la presunta existencia, para un determinado fenómeno, de una explicación psicológica que resultara irreductible a explicaciones neurofisiológicas, y otra distinta es que la descripción de los procesos neurofisiológicos en que en un determinado organismo o especie se encarnen los procesos psicológicos teorizados como explicación del fenómeno en cuestión sea irrelevante de cara a la construcción o a la justificación de la explicación psicológica. Es probable que nos veamos inclinados a pensar que de la autonomía de la psicología se sigue la irrelevancia de la fisiología precisamente si pasamos por alto las restricciones a la tesis de realizabilidad múltiple que encierra la cláusula condicional. Tanto la reconstrucción conceptual del razonamiento como un mínimo esbozo de su verosimilitud histórica quedan, sin embargo, forzosamente en el aire.

[208] El ejemplo de restricción de orden abstracto aducido por Block (1997: 121) es la posibilidad, a su juicio no desdeñable, de que las únicas estructuras que puedan abergar pensamiento sean o sistemas conexionistas o sistemas simbólicos clásicos.

[209] En torno a la distinción entre un nivel de diseño y uno de implementación, que es una de las vetas cuya extracción conduciría de por sí a una de las caracterizaciones más completas del cognitivismo, tanto en términos históricos como conceptuales, sería imprescindible contrastar, siquiera como punto

deberíamos esperar mayor homogeneidad en aquél que en este último. O dicho de otro modo, estas consideraciones apuntan, de nuevo, a que deberíamos esperar realizabilidad múltiple. La reivindicación de Putnam es ya inminente:

> Since evolution enforces similarity only at the design level, we should expect more variation at the levels of realization than at the design level. And this is why we expect multiple realization
>
> If there are distinct constraints at different levels, it is no surprise that as Putnam (1975) noted, different idealizations are appropriate at different levels. (Block 1997: 122)

Naturalmente, el envés de esa restitución de las intuiciones de Putnam es la contestación a los argumentos de Kim. Como ha quedado antedicho, cabe prever que las propiedades psicológicas a las que subyaga una mera peculiaridad de su particular encarnación física no muestren rasgos comunes entre diferentes implementaciones que convenga describir en términos funcionalistas; por el contrario, cabe prever que aquellas propiedades psicológicas cuya implementación física venga condicionada, además de por las restricciones que imponen las leyes naturales, por fuerzas como la evolución o el diseño, muestren de hecho patrones similares, desde un punto de vista funcional, en sus diversas encarnaciones. Las propiedades de instanciación –las del primer tipo– dan una parte de razón a Kim, mientras que las propiedades *D* –las del segundo tipo– atestiguan a favor de la concepción funcionalista de lo mental y, en particular, de la vigencia de la lectura antirreduccionista de la tesis de realizabilidad múltiple. Más a ras del suelo empírico: fenómenos como la aerodontalgia (la evocación de dolores dentales pasados provocada por la estimulación de la mucosa nasal) cuentan en el haber de Kim; fenómenos como la relación entre la ansiedad y el dolor agudo de aparición aleatoria o entre el dolor y la distracción se suman, a todas luces, en las cuentas funcionalistas. Aun ciñéndonos sólo al ámbito del dolor, parece claro que la concesión a Kim es, después de todo, de muy menor cuantía. O –tal como el propio Block (1997: 108) cifra sus conclusiones: "[…] much –but not all– of the anti-reductionist consensus survives Kim's critique".

Así guarnecido, Block retoma la ya ajada cuestión del jade. No debemos, a su entender, dar por buena la premisa de Kim según la cual "El jade es verde" no es un enunciado proyectable –o la propiedad de ser verde no es proyectable respecto del jade, si se prefiere. Puesto que el concepto de jade se define –piensa Block– en términos de apariencias, y puesto que el predicado"es verde" se refiere precisamente a dichas apariencias, cierto grado de proyección del predicado, cuando se aplica al concepto de jade, entra dentro de lo esperable[210]. No hay muchos motivos, en cambio, para esperar que las diversas muestras de jade –jadeíta o nefrita– compartan otras

---

de partida, sus avatares en Miller, Galanter y Pribram (1960, *infra*), Dennett (1973, 1981, 1987, *infra*), Newell y Simon (1975, *infra*), Marr (1982, *infra*), y Newell (1982; 1986, *infra*).

[210] Aunque, conviene apuntar, hay nefrita blanca –especialmente apreciada en la cultura china, donde se conoce como jade de grasa de cordero–, así como variedades de jadeíta amarillenta, rosada, violácea o azulada.

propiedades –"científicamente profundas", dice Block– más allá de las restricciones, seguramente más bien laxas, que el Principio de Disney imponga sobre las posibles configuraciones físicas de algo que haya de tener la apariencia del jade. Una elucidación parecida vendría al caso respecto del efecto sedante de distintos fármacos. Así que, de nuevo, algo de razón asiste a Kim, pero no tanta como sería preciso para sostener su impugnación del funcionalismo.

Dos lecciones de mayor envergadura cabe extraer, de la mano de Block, de tan espinosa discusión, así como una tercera de índole más tentativa. En primer lugar, debemos recordar que una propiedad sea o no proyectable depende de aquello sobre lo que la prediquemos. Releamos: *el enunciado* "El jade es verde" es, entonces, un enunciado proyectable, es decir, la propiedad de ser verde es proyectable *respecto del* jade –pero no hay una respuesta para la pregunta de si la propiedad de ser verde es proyectable en términos absolutos. Pero también: *cierto grado* de proyección del predicado, cuando se aplica al concepto de jade, entra dentro de lo esperable; no hay *muchos* motivos, en cambio, para esperar que las diversas muestras de jade –jadeíta o nefrita– compartan propiedades más básicas. Es decir, en segundo lugar: que una propiedad sea o no proyectable es una cuestión de grado, como lo es la similitud y, en consecuencia, la pertenencia a una clase o la inclusión en el ámbito de aplicación de un concepto[211]. En cuanto a la naturaleza del dolor:

> The upshot is that if pain is nomically equivalent to a physico-chemical disjunction, then both pain and the disjunction will be kinds with respect to some properties, but to a lesser degree with respect to others. Kinds are relative and graded. (Block 1997: 128)

La tercera conclusión, que queda en el aire cerrando las reflexiones de Block, es que el carácter gradual y relativo de la pertenencia a clases y del carácter proyectable de un enunciado nos fuerza a desligar ambas nociones –la de clase y la de propiedad proyectable– de la de causalidad, "[…] a no ser que uno esté preparado para ver la causalidad como algo relativo y gradual" (Block 1997: 129). A lo que parece, el propio Block no está por la labor de dar ese paso, pues se pregunta, retóricamente, cómo podríamos entender siquiera la tesis de que la causalidad es gradual.

---

[211] También son dos, cuando menos, los ecos que estas conclusiones despiertan de inmediato. Primero, vienen a la memoria –aunque Block no lo menciona– los trabajos seminales de Rosch (1973, 1975, 1978), y luego de Barsalou (1987) y Rips (1989) sobre conceptualización y categorización, en los que ya hemos visto confluir a la reflexión filosófica sobre el propio concepto de lo psicológico –por ejemplo, en Stich (1992: 352, *supra*)–, o sobre la delimitación de las diversas perspectivas teóricas –como en O'Donohue y Kitchener (1999: xx, *supra*). Segundo, resuena con fuerza –el propio Block lo anota– la advertencia de Davidson (1966) de que cualquier predicado es proyectable si es con respecto a la propiedad adecuada: remedando a Goodman (1953), hasta "verdul" o "verdojo" lo son –si no respecto a la propiedad de ser una "esmeralda", un "zafiro" o una "rosa", sí respecto a la de ser un "esmeriro" o una "esmerrosa", siempre que pongamos cuidado en que la definición de "esmeriro" cuadre con la de "verdul" y la de "esmerrosa" con la de "verdojo". Después de todo, la posición de Davidson ante Goodman parece casi de sentido común –si no fuera, claro, por lo lejos del sentido común que nos ha llevado la discusión–: a grandes males, grandes remedios.

Ha quedado en el camino, entretanto, un asunto que Block esquiva cuidadosamente en el debate con Kim, pero que resulta crucial en otros ámbitos. La cuestión adquiere ya un tono reiterativo: si Kim estuviera en lo cierto y no existiera al fin y al cabo *una* teoría cientifífica del dolor, sino tantas como mecanismos neurofisiológicos lo encarnaran, ¿qué justificación tendríamos para considerar elementos de la misma clase –es decir, casos de dolor– a aquellos cuya instanciación física difiriese entre sí? La ausencia de una respuesta abocaría a Kim al eliminacionismo; la respuesta que, según la interpretación de Block, habría ensayado Kim –a saber, que el concepto de dolor apunta a una propiedad de segundo orden: la de poseer alguna propiedad de primer orden que desempeña cierto papel causal– es tanto como un contundente refrendo del funcionalismo. Se trataría, ahora bien, de una variedad de funcionalismo que tanto Block como Fodor han recusado sin paliativos: no es, a su entender, el análisis conceptual del vocabulario psicológico ordinario lo que ha de develar las clases naturales que conforman lo mental, sino la investigación empírica en psicología. La vertiente del funcionalismo con la que Kim acaba por confluir, así pues, no es la que proviene de los fértiles aluviones de la teoría de autómatas que en su día labrara Putnam, o de la simulación computacional de procesos cognitivos que había comenzado ya a cultivarse tan pronto como la tecnología lo hizo factible, sino más bien la que, ingeniosamente canalizada por Smart, Armstrong o Jackson, desciende del trabajo de Ryle para regar las riberas de la teoría de identidad psicofísica y las vecinas huertas de la neurociencia cognitiva, con cuyos frutos alimenta Kim su argumentación[212]. Pero conviene posponer el examen de esta cartografía del pensamiento funcionalista hasta dar por concluido, siquiera momentáneamente, el diálogo sobre la presunta heterogeneidad del sustrato físico de lo mental, y sobre su interpretación tanto en términos epistemológicos como ontológicos.

Ni la respuesta de Block (1997) ni la de Fodor (1997) convencen a Jacob (2002), quien carga las tintas sobre la idea de que el propio razonamiento de Kim lo aboca sin remedio al eliminacionismo, con consecuencias autorrefutatorias. En realidad, como bien señala Jacob (2002: 649), el compromiso inicial de Kim con el realismo acerca de lo mental se perfila a la vez como un punto de concordancia con los planteamientos funcionalistas, de cuya lectura antirreducionista Kim discrepa, y como uno de discordancia con otra de las principales vertientes que el pensamiento antirreducionista presenta en el ámbito del naturalismo, y que Kim se propone descabalgar (*cf*. *infra*): el monismo anómalo de Davidson. La cuestión es, en suma, que Kim no admite que lo mental sea autónomo, con Davidson, por razón de su anómala sujeción a un entramado normativo, ni tampoco que lo sea, con la ortodoxia funcionalista inspirada en Putnam, por razón de su variable materialización: dicha

[212] También en el trabajo de Armstrong y Lewis encuentra Kim (2000: 89-102; 2002: 642) las fuentes de su idea de reducción funcional, con la que trata de reemplazar el modelo nageliano de reducción mediante leyes-puente de modo que el argumento antirreduccionista de Putnam y Fodor, que acomete contra dicho modelo, quede en el aire.

variabilidad, que Kim asume, excluye de hecho, a su entender, la noción de autonomía –antes al contrario, el funcionalismo entraña reduccionismo[213].

Pues bien, la cardinal debilidad de la argumentación de Kim residiría de acuerdo con Jacob precisamente en su réplica a la pregunta que Block ha hecho suya acaso con más pertinacia que otros, y que Lewis (1969: 233, *infra*) despreciara como fruto de viejos prejuicios: ¿en virtud de qué cabe tomar como elementos de la misma clase psicológica a aquellos cuya instanciación física resulte heterogénea? No obstante, Jacob llama la atención sobre un aspecto de la respuesta de Kim que habría pasado inadvertido a ojos de Block: "[…] Kim entertains the radical view that one ought to trade higher-order functional properties for higher-order functional concepts" (Jacob 2002: 653). Pero, por supuesto, es difícil ver "[…] what the relevant notion of a second-order concept might be other than the concept of a second-order physical property". Ésa –cómo no– es la reconstrucción de la propuesta de Kim que opera en el propio Block, quien, como recién se ha visto, emplaza a Kim en el entorno del funcionalismo analítico de Lewis: así, los conceptos psicológicos apuntarían a *propiedades* de segundo orden consistentes en la posesión de propiedades de primer orden que desempeñen determinado papel causal. La objeción de Jacob a la maniobra amagada por Kim es tajante: si los conceptos psicológicos no son conceptos de propiedades físicas de orden superior, es decir, de propiedades funcionales, lo único que puede retraer a Kim del eliminacionismo es que esté dispuesto a concluir que son conceptos de propiedades no instanciadas. Comoquiera que tal conclusión es inaceptable, la propuesta de Kim –asegura Jacob (2002: 653)– equivale a una renuncia al reduccionismo y una adhesión al eliminacionismo. Pero llegado tal caso, en un pulcro quiebro del debate, toda la argumentación erigida por Kim se desmorona súbitamente al contradecir su conclusión uno de los presupuestos sobre los que se cimentó, a saber, el realismo psicológico. En efecto:

> […G]iving up mental properties suffices to remove the question: if some physical property of a cause was efficacious in the process whereby it produced its effect, then could its mental property be efficacious too? If so, then the problem of causal exclusion simply does not arise and Kim's supervenience argument is undermined. (Jacob 2002: 654)

Lo cierto es que Kim (2002: 462-643) parece dar a este asunto un lugar primordial en su revisión de las preguntas que quedan pendientes a la luz de su propia propuesta. Aun cuando todos las instancias *Mi…Mn* de una propiedad mental *M* hayan

[213] Aunque, como también apunta Jacob (2002: 649), Kim parece deponer su ánimo reduccionista cuando se trata no ya de la intencionalidad de lo mental sino de su carácter fenoménico o cualitativo, que bien podría resultar –concede– refractario al análisis funcionalista y la reducción neurofisiológica: "[…] perhaps, we must learn to live with qualia epiphenomenalism" (Kim 2002: 643). De esta manera, el empuje de Kim se suma a la propensión a desligar el estudio de la intencionalidad y el de la consciencia que, pese a las protestas de McGinn (1988: 299) –quien dio en bautizar dicha propensión como "the insulation strategy"– o de Searle (1992: 153), ha sido, como quedó antedicho, poco menos que hegemónica en el seno del cognitivismo; *cf.*, nuevamente, Chacón y Hermoso (2009).

quedado satisfactoriamente identificadas con sus respectivos trasuntos físicos *Fi…F*n*, presumiblemente heterogéneos, cabrá preguntarse si la propiedad *M* –es decir, si el tipo de estado mental delimitado por *M*– habrá quedado *ipso facto* reducida a *F*∨*F**. Cara a cara con la cuestión, Kim (2002) tantea la postura de Lewis (1969) –impugnarla–, y a renglón seguido tanto la conclusión eliminacionista a la que Jacob (2002) lo ve abocado como la confluencia con el funcionalismo analítico en la que lo ubicara Block (1997) –donde, desde luego, parece encontrar un paraje más grato, pese a la necesidad de convivir con propiedades físicas de naturaleza indefinidamente disyuntiva.

> All we need to worry about in the matter of M's reduction, I believe, is the reduction of all M-instances; there are no significant philosophical issues beyond that. However, if one is concerned with the ontological fate of M, there are two possibilities. The first option is this: we deny that there are such things as "second-order" properties (note: functional properties are standardly construed as a species of second-order properties); there are only second-order predicates, expressions, and concepts. By forming second-order expressions we do not bring into existence new properties; we only introduce new ways of talking about the properties already in hand. This is an irrealist, a sort of eliminativist, option. Second, there is a conservative (preservative or retentive, as some would put it) option: Identify M with the disjunction of its realizers. The disjunction will have indefinitely many disjuncts, but I don't see that as a problem. We of course need an ontology that allows such disjunctions, but I don't see a deep problem here either. On this approach, M will be retained as a disjunctive property –causally and nomologically heterogeneous and unprojectible, but a property nonetheless. (Kim 2002: 642-643)

Respecto a que esto constituya una ratificación de la variedad de funcionalismo que han abanderado Lewis, Armstrong o Jackson, es importante consignar una salvedad: si bien es cierto que el primer paso del procedimiento de reducción funcional delineado por Kim pasa por que la propiedad reducida quede "[…] reformulada como una propiedad de segundo orden por medio de una defición […] en términos de su "papel causal" […]" (Kim 2002: 642) –entrecomillar "papel causal" bien parece, por parte de Kim, un reconocimiento de esas deudas–, no es menos cierto que ese primer punto del expediente reductivo se describe como "[…] un paso conceptual, aunque […] propenso a reflejar numerosas restricciones y desiderátums teórico-empíricos". No es arriesgado entrever aquí un intento de rebajar la discrepancia entre una interpretación analítica y una empírica del funcionalismo, al menos en lo concerniente a la vía –conceptual o empírica– por la que hayamos de procurarnos el análisis de los conceptos psicológicos, convirtiéndola en una cuestión de proporciones.

Por lo demás, como veíamos en relación con los movimientos de retirada orquestados por Block (1997: 115, *supra*) ante la ofensiva de Kim, no resulta impensable que la ruta abierta por Lewis (1969), por la que Kim intenta soslayar la liquidación eliminacionista de lo mental, acabe por conducir de todos modos a ese mismo desenlace. Cuando menos, conviene insistir en que Kim parece abandonarse a su suerte, o acaso a su intuición fisicalista, en cuanto a que la realizabilidad múltiple

de los estados psicológicos se detendrá antes de llegar a un nivel irremediablemente idiográfico, pero no nos proporciona razones que hagan pensar que tal cosa es improbable.

Distintas mañas pero el mismo propósito que en Fodor (1997) –o sea, hallar suelo firme sobre el que dar por zanjado el intercambio de intuiciones contrarias– puede entreverse en el intento de Witmer (2003) de concertar como una premisa eximida de la controversia la de que las propiedades a las que se refieren nuestros conceptos psicológicos forman, en las condiciones adecuadas, predicados proyectables. Esto vendría a ser para Witmer (2003: 67) un "hecho mooreano", en referencia al afamado ensayo de refutación del escepticismo sobre la existencia de un mundo exterior a uno mismo protagonizado por George E. Moore en 1939 al argumentar que actos como el de levantar una mano y decir "He aquí una mano", luego otra y decir "He aquí otra mano", son razón suficiente para descreer de las intuiciones filosóficas que daban aliento a convicciones como las de su otrora profesor John McTaggart respecto a la irrealidad del tiempo (McTaggart 1908). En efecto, si hemos de asumir como un principio del sentido común que ciertas generalizaciones formadas mediante el uso de conceptos psicológicos son predicados proyectables, entonces el dilema planteado por Kim se vence irremisiblemente –salvo que podamos evitarlo con argumentos como los que esgrimen Fodor (1997, *supra*) o Block (1997, *supra*)– del lado que el propio Kim parece preferir: puesto que la propiedad psicológica en cuestión también lo es, la disyunción de propiedades fisiológicas que corresponde a una propiedad psicológica es proyectable. Así pues, sería un reduccionismo a la vieja usanza, seguramente cercano al espíritu de Lewis (1969), lo que el razonamiento de Witmer avivaría, en detrimento de las conclusiones eliminacionistas que acaban por desmentir el declarado realismo psicológico de Kim. Ahora bien, dicho razonamiento da por buena la premisa de Kim en cuanto a que dos propiedades que determinen clases nomológicamente coextensivas han de ser ambas proyectables o no serlo ninguna, premisa cuya impugnación ya hemos recorrido de la mano de Fodor (1997), Block (1997) y, *avant la lettre*, Boyd (1980).

Es buena hora para sopesar cuánto queda en pie, entonces, del vigoroso argumento de Kim y de las objeciones planteadas. El examen del alegato erigido por Fodor arroja resultados dispares. Por una parte, parece sólida su reafirmación de las razones que nos hacen preferir generalizaciones universales antes que enumeraciones abiertas, aun a costa de incurrir en cierto grado de hipóstasis de las propiedades que sustentan la generalización. Advertir la imbricación de esas razones en el mismo esqueleto de nuestras prácticas inductivas respalda la obstinación de Block en preguntar por aquello que hace de cada episodio de dolor un elemento de la misma clase de estados mentales, suponiendo que la encarnación fisiológica de distintos episodios resulte irreconciliablemente dispar, a la vez que contrarresta la displicente derogación de esa pregunta que Lewis (1969: 233, *infra*) intenta al asemejarla a viejos prejuicios metafísicos, así como el intento de Kim de hermanar funcionalismo y reduccionismo en el rechazo de la idea de que cada estado psicológico haya de poseer una naturaleza intrínseca. Además, la posición de Fodor a

este respecto enlaza con el concienzudo trabajo que Pylyshyn (1984: 4, *infra*) dedica a fraguar la idea de que distintos vocabularios proporcionan distintas generalizaciones dejándola reposar sobre la distinción entre la extensionalidad de la descripción y la intensionalidad de la explicación –lo cual resulta natural, por supuesto, considerando que el propio Pylyshyn había cobijado su análisis bajo la inspiración de Fodor y Block (1972b). Una buena señal: en ese mismo estuario parecen confluir también, desde fuentes opuestas en su valoración del reduccionismo, las advertencias de Sober (1999: 550, *supra*) en cuanto a que la minuciosidad a la hora de rendir cuenta de los procesos involucrados en un suceso particular es un objetivo tan digno de anhelo en el seno de la labor científica como la generalidad a la hora de envolver muchos de esos sucesos en un enunciado legaliforme, y la benévola relectura de Putnam (1975b) que en ese sentido elabora Block (1997).

Por otra parte, sin embargo, el esfuerzo de Fodor por quebrar la analogía entre el dolor y el jade, a fin de sortear las conclusiones que de ella pretende extraer Kim, nunca parece llegar a buen puerto. Que la identidad del jade con la disyunción cerrada jadeíta-o-nefrita sea de índole necesaria queda pobremente establecido, y ni siquiera el ensayo de reformular la cuestión bajo un prisma lakatosiano conduce, en esos términos, a una refutación terminante del reduccionismo. Acaso la única escapatoria que quede para Fodor sea admitir que la totalidad de las muestras de jade constituya al fin y al cabo una clase cuya realización física es variable en el sentido técnico relevante, y sobre la cual por tanto haya de ser posible formar enunciados proyectables irreductibles a los que puedan predicarse como tal de las diversas realizaciones fisicas del jade, sin que ello sea óbice para apuntar los innumerables matices que diferencian sin duda esa clase de la formada por todos los episodios de dolor. La ruta, a partir de ahí, pasaría por recabar las generalizaciones acerca del jade que quepa rescatar de la ciencia especial que hubiera de tomarlo como objeto –presumiblemente la gemología, tal vez incluso la economía–, hacer constar la autonomía de tales generalizaciones respecto de las que puedan formularse en el vocabulario teórico de la mineralogía, la química inorgánica o la física, y explicar las diferencias con respecto a la cuantía y la envergadura de las generalizaciones alcanzadas en el caso del jade y las que pudiera arrojar el caso del dolor como resultado precisamente de los aspectos que diferencian a ambas clases entre sí, entre los cuales ocuparía seguramente un lugar destacado el desparejo entrelazamiento de una y otra en los usos cotidianos y en nuestra visión del mundo.

Las primeras mieses de la cosecha que augura esa táctica se recogen al abordar el problema del jade. Recordemos: un científico que tratara de corroborar la hipótesis de que el jade es verde no obtendría justificación racional para hacerlo mediante la constatación de que todas las muestras recogidas son verdes, dado que bien podría darse la doble circunstancia de que todas ellas fueran muestras de jadeíta, y de que al menos algunas muestras de nefrita no fueran verdes; luego las propiedades cromáticas del jade no son proyectables, o nomológicas, y no lo son según parece en razón de la estructura disyuntiva de la clase en cuestión. El desafío de Kim a la interpretación antirreduccionista del funcionalismo se cifra en una pregunta simple:

¿por qué habría de ser diferente el caso de los estados mentales cuya encarnación física resulte ser variable? Tenemos por un lado la tajante respuesta de Fodor (1997): el caso del jade no llega siquiera a constituir un indicio fiable respecto de la naturaleza de los estados mentales, porque hay razones para pensar que la analogía es espuria. Una actitud algo más receptiva al razonamiento de Kim se deja ver en Block (1997): que tal o cual propiedad resulte proyectable no es un asunto dicotómico, sino de grado, y hay razones tanto para pensar que diferentes muestras de jade compartan algunas, pero no muchas, propiedades, como para pensar que diferentes episodios de dolor, o de otros estados mentales, exhiban –por así decir– una estructura metafísica con mayores y más hondas vetas comunes. Ahora bien, el propio análisis de Block contiene ya la conclusión de que "ser proyectable" es una propiedad –si se quiere, una metapropiedad– relativa, o una relación entre dos propiedades: ninguna propiedad es ni deja de ser proyectable en términos absolutos, sino únicamente en el seno de un enunciado, en función de aquello de lo que se predique. Así, "ser verde" parece ser proyectable en bajo grado respecto de "ser jade", pero –cabe añadir– seguramente lo sea en un grado mayor respecto de "estar formado de cloroplastos activos" [214]. La causa, además, es precisamente la que cabe esperar dada la distinción entre propiedades de instanciación y propiedades *D*, tensadas por restricciones físicas y evolutivas, en la que Block cimenta su respuesta a Kim: *grosso modo*, que no existe parangón en el ámbito de los inosilicatos, como la jadeíta y la nefrita, para la relación entre la función desempeñada por la clorofila en la fotosíntesis oxigénica y la curva de absorción de radiación electromagnética en el espectro visible que exhiben las moléculas de clorofila[215].

[214] Es discutible, por supuesto, que "estar formado por cloroplastos activos" sea además una propiedad cuya realizabilidad física sea variable. Aunque existen al menos cinco tipos de clorofila cuyas estructuras moleculares son de hecho levemente diferentes, no parece descartable en principio que las diferencias sean asumibles con los recursos teóricos propios de la química orgánica, sin necesidad de acudir a un vocabulario explicativo de orden superior, en el seno de la fitología, para dar cuenta del hecho de que todas esas moléculas sean clorofilas. Por otra parte, existen también pigmentos rojos, pardos, amarillos e incluso azules con actividad fotosintética, como la ficoeritrina y los carotenoides, y tres de las variedades de clorofila están presentes en algas pardas, como el kelp, en algas rojas, o en cromoalveolados de morfología muy diversa, como las algas diatomeas, que producen de hecho buena parte del oxígeno atmosférico. Nada de ello entorpece el argumento: no se ha tratado de establecer que "ser verde" resulte plenamente proyectable respecto de "estar formado por cloroplastos activos" –es decir, que proporcione plena justificación para suponer que cualquier cosa formada por cloroplastos activos que podamos encontrar haya de ser verde–, sino que resulta proyectable en mayor medida que respecto de "ser jade"; tampoco se ha planteado que una misma propiedad pueda ser más o menos proyectable respecto de otras siempre que la instanciación física de esas otras sea variable, sino respecto de otras, sin más.

[215] En otro orden de cosas, conviene dejar anotado que el hecho de que la proyección del predicado "ser verde" a algo de lo que sabemos que está formado por cloroplastos activos pudiera estar más justificada que la proyección del mismo predicado a algo de lo que sabemos que es jade resulta en cierta medida independiente de las consideraciones de aire duhemiano o lakatosiano que se han venido articulando. Es indudable que el verdor de las plantas es algo a lo que seríamos más reacios a renunciar que el verdor del jade, o, visto de otro modo, que un mundo en el que las plantas no fueran verdes nos resultaría mucho más extraño que uno en que no lo fuera el jade; en ese sentido, el hecho

Bien es posible, en cambio, que existan predicados formulados en un vocabulario explicativo adecuado –el de la gemología, como decíamos– y cuya proyección respecto del jade resulte mucho más robusta que la de "ser verde", como por ejemplo "El precio de mercado del jade es mayor que el de la serpentina", o sobre cualquiera de los inosilicatos que tomamos como jade, tal como "La jadeíta pura es más valiosa que la jadeíta cuya transparencia ha sido reforzada por impregnación con polímeros". En efecto, si tenemos entre manos una gema de jadeíta pura, no es aventurado inducir que su precio de mercado será superior al de la jadeíta estabilizada, al igual que lo ha sido en las muestras anteriores, y también, desde luego, al de la serpentina. Del mismo modo, si hemos de dar con predicados que resulten proyectables con respecto a determinados estados psicológicos, es de esperar que se encuentren saturados del vocabulario teórico de la psicología. Los ejemplos aducidos por Block, como que el dolor agudo de aparición aleatoria causa respuestas de ansiedad (Block 197b: 117), o que en los procesos de aprendizaje, tal como ha mostrado Shepard (1987 *apud* Block 1997: 124-125), la probabilidad de generalización a un estímulo se ajusta a una función exponencial decreciente de la distancia en un espacio psicológico abstracto ya de métrica euclidiana ya ortogonal, no dejan lugar a dudas en este sentido.

Diferentes vocabularios teóricos, en fin, se aprestan a diferentes generalizaciones –ya se ha recordado el trabajo de Pylyshyn (1984) al respecto, sobre el que habrá ocasión de regresar–, y diferentes generalizaciones muestran a su vez diferentes grados de acato a la proyección de las propiedades que hipostatizan. Que tal cosa sea posible –que existan clases definidas en vocabularios explicativos distintos del de la física y cuya condición de clase, como quiere Fodor, venga avalada por su presencia en generalizaciones verdaderas y no triviales distintas de las propias leyes-puente que las engranan con sus implementaciones físicas– y que todo cuanto existe sea en último término de naturaleza física son –qué duda cabe– dos ideas que nos resulta difícil reconciliar. Esa tensión parece haber llevado a Kim a construir un argumento que trata de refutar la posible existencia de vocabularios teóricos explicativamente autónomos –es decir, de ciencias especiales– partiendo de principios, como el de clausura causal de lo físico o el de herencia causal que no resulten controvertidos salvo para quien rechace de antemano la naturaleza física de lo real. De acuerdo con Fodor (1997), la propuesta de Kim conduciría de hecho a un

---

de que las plantas tienen partes verdes –aunque los legos acaso no las identifiquemos como órganos vegetativos autótrofos– parece alojarse más cerca del corazón de nuestras creencias sobre el mundo que ningún hecho relativo al jade. Pero no parece que opusiéramos mayor reticencia a admitir como plantas a organismos que careciesen de órganos vegetativos de color verde que la que tenemos en admitir que una gema de color rosado, violáceo, amarillento, o incluso blanco, sea jade. Tampoco cambian mucho las cosas al pasar del conocimiento cotidiano al terreno de la ciencia: de hecho, las fronteras del reino *Plantae* alientan un reñido debate taxonómico en el que se plantea, entre otras, la posibilidad de que el concepto de planta deba ser eliminado –aunque nadie parece deducir de ello que *en realidad* no existan las plantas–, y también la de que deba hacerse corresponder precisamente al clado *Viridiplantae*, formado por las plantas terrestres o embriofitas y las algas verdes, que heredaron sus cloroplastos de un antecesor común con las algas rojas y las glaucofitas.

alivio de la tensión, en la medida en que sólo subsistirían aquellas generalizaciones a las que subyacieran regularidades derivadas de leyes físicas que, a su vez, expresaran relaciones entre clases físicamente homogéneas. Lo malo –parece pensar Fodor– es que la fuerza de dicha propuesta se agota en el hecho de que promete una conclusión tranquilizadora; por lo demás, no va más allá de lo que iría una mera estipulación:

> Kim wants to *just stipulate* that the only kinds there are are (what he calls) local; viz. the only kinds there are are the kinds of kinds whose realizers are physically homogeneous. (Fodor 1997: 161)

Establecer la conclusión que Kim pretende resultaría incomparablemente más fatigoso, como Fodor con deliberada prolijidad se esmera en hacer patente:

> Science postulates the kinds that it needs in order to formulate the most powerful generalizations that its evidence will support. If you want to attack the kinds, you have to attack the generalizations. If you want to attack the generalizations, you have to attack the evidence that confirms them. If you want to attack the evidence that confirms them, you have to show that the predictions that the generalizations entail don't come out true. If you want to show that the predictions that the generalizations entail don't come out true, you have actually to *do the science*. Merely complaining that the generalizations that the evidence support imply a philosophically inconvenient taxonomy of kinds *cuts no ice at all*. (Fodor 1997: 161-162)

Aunque el juicio de Fodor seguramente sea demasiado severo, sus reparos se cuentan entre los que hacen que el argumento de Kim no pueda tomarse como concluyente. Es obvio, sin embargo, que la tensión que motivó dicho argumento sobrevive a su descalabro.

Sea como sea, parece claro que buena parte de la defensa del funcionalismo ante el renovado brío reduccionista transita con naturalidad entre la realizabilidad múltiple, que es una cuestión ontológica, y la relevancia explicativa de determinadas estrategias taxonómicas, asunto de índole epistemológica. Pero hemos visto ya, en relación con los argumentos de Braddon-Mitchell y Jackson (1996, *supra*), que la noción de relevancia explicativa no permite más que una reconstrucción muy deficitaria de la idea de autonomía explicativa que parece operar en el seno del cognitivismo, la cual bien podría exigir que halláramos el resquicio que permitiese dotar a lo mental de eficacia causal propia. Ahora bien: como acaso con mayor precisión que otros ha hecho notar Liz (1995), cuando hemos de llegar desde la relevancia explicativa hasta la eficacia causal el camino no se muestra tan despejado:

> Resulta difícil imaginar cómo podría llegar a ser razonable admitir la eficacia causal de lo mental si la apelación a esos fenómenos mentales no pudiera ser *nunca* explicativamente relevante. Pero no lo es tanto imaginar una relevancia explicativa de lo mental *sin* que lo mental pudiera llegar a ser causalmente eficaz. La apelación a fenómenos mentales puede ponernos sobre la pista de ciertas relaciones causales en las que, sin embargo, esos fenómenos mentales no intervengan. (Liz 1995: 212-213)

Lo que tendríamos en tales casos es una generalización –más exactamente, una ley *caeteris paribus*– que resultaría explicativamente relevante merced al hecho de que "[…] recoge cierta regularidad sin contrapartida directa con ninguna relación causal que responda a los términos de esa regularidad" (Liz 1995: 213). Por supuesto, la cláusula *caeteris paribus* tiene el cometido, en este tipo de enunciados, de mantener constantes las "[…] complejas relaciones causales subyacentes que explican esa regularidad" (Liz 1995: *íbid*.)

Para sumar eficacia causal a la relevancia explicativa, en cambio, tendríamos que hacer patente la existencia de "[…] genuinas relaciones causales que, bajo las restricciones impuestas por la cláusula *caeteris paribus*, son *detectadas* por los términos en los que expresamos nuestra ley" (Liz 1995: 214). La dificultad radica, entonces, en reconciliar la idea de que existen relaciones causales genuinas entre fenómenos psicológicos con el principio general de clausura causal de lo físico que es medular a una concepción naturalista del mundo. Dicho de otro modo: si cada instancia de un fenómeno mental es ni más ni menos que una instancia de un fénomeno físico, parece medianamente comprensible que, en virtud de su clasificación bajo un tipo de fenómenos mentales que no viene articulado según criterios físicalistas, *nosotros* atesoremos alguna ventaja explicativa; así sucedería, en efecto, si nuestra clasificación hubiera topado con una veta de relaciones causales físicas y pudiera rastrearla gracias a las cláusulas *caeteris paribus*. Pero, si cada instancia de un fenómeno mental es ni más ni menos que una instancia de un fénomeno físico, no parece ni siquiera medianamente comprensible –al menos a primera vista– que, en virtud de su pertenencia a un tipo de fenómenos mentales que no viene articulado según criterios físicalistas –ni, de hecho, en virtud de ninguna otra cosa que no vulnere la identidad postulada–, *el fenómeno mental* adquiera alguna relación causal que no poseyera igualmente en tanto que fenómeno físico. Este es, al fin y al cabo, el principio de herencia causal que Kim (1993a: 355, *supra*) atribuye a la relación de instanciación, o realización.

Hay –es cierto– una asimetría entre los requisitos que nos permitirían establecer la relevancia explicativa de una generalización en la que intervienen términos teóricos referidos a estados mentales y los que avalarían la eficacia causal de tales estados mentales. Pero la asimetría, como vimos, se disipa si reemplazamos el desiderátum de relevancia explicativa por el de irreductibilidad. Para que pudiéramos concluir que un enunciado científico legaliforme –una ley *caeteris paribus,* o, si se prefiere, la conjunción de las generalizaciones recogidas en un modelo teórico– es irreductible y, en consecuencia, que lo es al menos uno de los conceptos teóricos que emplea, el enunciado tendría que capturar regularidades que no fuera dable apresar, ni expresar, mediante los recursos conceptuales de las ciencias más básicas –en último término, de la física. Ahora bien –mimetizando el giro de Liz (1995: 212-213, *supra*)–, resulta difícil imaginar una irreductibilidad de lo mental *sin* que lo mental pudiera llegar a ser causalmente eficaz, sin que los estados mentales alcanzaran a trabar las tan traídas y llevadas relaciones causales genuinas *como tales*

estados mentales. El caso es que la de irreductibilidad, como la de relevancia explicativa, es una noción epistemológica. Esto muestra que la asimetría no proviene del hecho de que la cuestión de la relevancia explicativa sea de orden epistemológico y la cuestión de la eficacia causal de orden ontológico, sino más bien de la excesiva laxitud del concepto de *relevancia*, que tan perniciosa resultara para los argumentos contra la realizabilidad múltiple pergeñados por Shapiro (2000, 2004). Dado que la polémica en torno a la posibilidad de variaciones en la encarnación de los fenómenos mentales atañe ante todo a las consecuencias antirreducionistas que de ello se ha visto desprenderse, seguramente agilicemos la discusión si convenimos en dejar de lado provisoriamente la cuestión de la relevancia explicativa. Entiéndase: no es que establecer la relevancia de la explicación psicológica fuese –después de tanto– irrelevante, sino que de cara a la cuestión de la eficacia causal de lo mental, y con ello del realismo respecto de los fenómenos mentales, sería una conclusión inerte; lo más lejos que podría llevarnos es acaso hasta un dúctil instrumentalismo.

En la misma línea apunta Jacob (2002: 651) que los ensayos de cimentar la eficacia causal de propiedades de orden superior sobre su participación en generalizaciones verdaderas, como los que atribuye a Block (1997) o a Antony y Levine (1997), franquean indebidamente el paso "de la causalidad a la explicación". Pero la cuestión no es sólo que la de explicación sea una noción epistémica donde las halla y la de causalidad una noción ontológica, como parece suponer Jacob: igualmente epistémica es sin duda la noción de autonomía explicativa, entendida como irreductibilidad, que Jacob no diferencia expresamente de la de relevancia, y, en cambio, la irreductibilidad de la explicación psicológica sí parece abrigar repercusiones inmediatas en el terreno de la causalidad, y viceversa. Parece, así pues, que es menester para Jacob cuestionar con mayor detenimiento su convicción, compartida con Kim, de que "[…a]ny attempt to overcome this problem by distinguishing different kinds of explanation will fail to engage with the metaphysical issue generated by the metaphysical assumption that mental properties are irreducible" (Jacob 2002: 651) –una idea que, irónicamente, Jacob confiesa no haber llegado a entender hasta hace algún tiempo.

También Fodor (1997), en su convencida defensa de la autonomía de las ciencias especiales frente a la arremetida aguijada por Kim (1992), parece verse en el mismo callejón. Desde los primeros renglones del trabajo queda claro que de lo que se trata es de dilucidar "[…] whether functional states are *ipso facto* autonomous" (Fodor 1997: 149) –se entiende, por el contexto, que la cuestión es obviamente aclarar si son *ipso facto* autónomas las leyes o teorías en las que se mencionan y especifican dichos estados funcionales, no los propios estados. Como se ha detallado, la conclusión de Fodor es afirmativa: a diferencia de las propiedades meramente disyuntivas, como "ser una muestra de jade", que forman clases cerradas y predicados no proyectables, y que en consecuencia no ameritan figurar en generalizaciones nomológicas, las propiedades cuya realización es variable, como sería el caso de las propiedades psicológicas, forman clases abiertas y predicados proyectables, de modo que bien pueden figurar –como de hecho ocurre, según

estima Fodor– en generalizaciones nomológicas que no son susceptibles de reducción a generalizaciones sobre la disyunción abierta de propiedades físicas en que se encarnan, y que por tanto acrecientan el erario explicativo de las ciencias especiales. Ahora bien, llegada la hora de dar cuenta de las razones por las que una descripción articulada según propiedades funcionales de nivel superior, como las psicológicas, es capaz de capturar generalizaciones que eluden a una descripción articulada según una disyunción de propiedades más elementales, incluso siendo así que siempre que se instancia la propiedad funcional en cuestión se instancia también alguno de los términos de la disyunción, esto es, llegada la hora de rendir cuenta de la autonomía de la explicación psicológica en términos de la eficacia causal de lo mental, el propio Fodor (1997: 159, *supra*) se declara –recordemos– doblemente impotente: ni sabe cómo hacerlo ni cómo, siquiera, *pensar sobre cómo hacerlo*[216]. Esta perplejidad, en todo caso, no atañe ya a la eficacia causal de lo mental, sino que impregna –o eso piensa Fodor– toda regularidad naturalmente observable. Así que no nos queda más remedio, nos guste o no, que "aprender a vivir" (Fodor 1997: 162) con el hecho de que:

> Damn near everything we know about the world suggests that unimaginably complicated to-ings and fro-ings of bits and pieces at the extreme *micro*level manage somehow to converge on stable *macrolevel* properties. (Fodor 1997: 160)

De concretar cómo sucede tal cosa es, en definitiva, de lo que Fodor desiste, si bien su propia postración le sirve, como vimos, para ofrecer un diagnóstico de la –legítima, aunque descarriada– inquietud filosófica que lleva a Kim a intentar desestimar la proyectabilidad de las propiedades que no sean físicamente homogéneas. La tarea pendiente es, por tanto, la de reconciliar el nudo materialismo y la eficacia causal de las propiedades funcionales, y Kim habría acertado, al menos, en señalar que esa tarea está pendiente.

Dotar a lo mental de una eficacia causal que le sea inalienable se perfila, a la luz de todo esto, como el reto mayor que se le plantea al funcionalismo, al menos en su interpretación antirreduccionista. No en vano, ni siquiera está claro que pueda mantenerse en pie un realismo sobre lo mental mínimamente consistente si no es sobre la reivindicación de alguna noción de eficacia causal que resulte aplicable a

[216] Acaso para disipar la impresión de que pudiera estar cediendo al *¡ignoramus!* de du Bois-Reymond, o incluso acaso a su *¡ignorabimus!* –como, por otra parte, no tuvo reparo en hacer respecto a la psicología de los procesos centrales en su "Primera Ley de la Inexistencia de la Ciencia Cognitiva" (Fodor 1983: 107): "[…] the more global […] a cognitive process is, the less anybody understands it"–, o tal vez sólo por la fuerza del hábito, el tono adoptado por Fodor es abiertamente jocoso, muy lejos de la solemnidad decimonónica de du Bois. El reconocimiento de su ignorancia resulta –nos dice Fodor (1997: 161)– "[…] embarrassing for a profesional philosopher –a paid up member of the A[merican] P[hilosophical] A[ssociation], Eastern Division", pero cuenta con superar dicha ignorancia "[…] the day before I figure out why there is anything at all, another (and, presumably, related) metaphysical conundrum that I find perplexing". No ya du Bois, pues, sino el mismísimo Leibniz.

nuestras creencias, deseos o aflicciones en tanto que determinantes de nuestro comportamiento. Como bien apunta Liz (2012: 75):

> Sin ninguna eficacia causal, por ejemplo respecto a la acción, no parece haber demasiadas razones para considerar que realmente exista algo que propiamente pueda caracterizarse como mental.

Pues bien, la táctica tal vez más común para afrontar dicho reto apela al contenido de los estados intencionales, de modo que el problema de la eficacia causal de los estados mentales se condensa en el de la eficacia causal de su semántica –y la autoridad de la concepción funcionalista de la mente recae, en gran parte, sobre la verosimilitud de su enfoque del contenido de creencias, deseos, intenciones o anhelos. Ésta es, en lo esencial, la misma conclusión a la que, por otros derroteros, conducían ya las reflexiones de Liz (1995): no habrá teoría cabal de la causalidad de lo mental mientras no la haya de la semántica de lo mental.

> […S]i tuviéramos una teoría naturalizada del contenido, la mayoría de los problemas que hemos discutido se resolverían por si sólos. Los análisis que se hagan de esa clase tan peculiar de explicaciones causales que mencionan fenómenos mentales son más directamente dependientes de nuestras concepciones semánticas que de nuestras concepciones acerca de la causalidad. (Liz 1995: 239)

Acaso entretanto no podamos hacer mucho más que, como Fodor (1997: 159), resignarnos jovialmente a convivir con ese aire "*molto mysterioso*":

> So, then, *why is there anything except physics*? […] Well, I admit that I don't know why. I don't even know how to *think about why*. (Fodor 1997: 161)

El instante oculto y múltiple en el que la latitud semántica de lo mental parece encerrar el misterio de ese carácter proteico de la realidad –casi como el libre albedrío se condensaba para Epicuro de Samos en la *parenklisis*, una levísima desviación espontánea, que Lucrecio (2.216-93) llamaría *clinamen*, de la trayectoria de ciertos átomos– parece siquiera vaguísisamente acotado:

> Lots of different sorts of micro-interactions manage, somehow or other, to converge on much the same macro-stabilities. The world, it seems, runs in parallel, at many levels of description. (Fodor 1997: 162)

Pero aun a sabiendas de que quizá no sea sino esa rendida perplejidad lo que aguarde al final del camino, es preciso transitar todavía algunos de sus recodos, tanto alrededor de la propia tesis de que varios estados psicológicamente homogéneos pueden encarnarse en otros físicamente heterogéneos –aunque tomados de uno en uno no sean otros sino el mismo– como en torno a la interpretación de las repercusiones epistemológicas y ontológicas de esa tesis, así como, particularmente, en torno al intento de entender de qué manera el hecho de que los estados mentales aludan a realidades distintas de ellos mismos pudiera abrigar las respuestas a, al menos, algunas de estas preguntas.

Prácticas de taxonomía neurológica y psicológica: estructura y función

Junto con los trabajos de Kim, la otra noticia que ha dado pábulo según Wilson y Craver (2007) al cuestionamiento de la tesis de realizabilidad múltiple viene dada, como ya se anticipó, por los vertiginosos avances que la neurociencia cognitiva ha logrado en tiempos recientes. Al hilo de esos avances, Bechtel y Mundale (1999) han erguido un ataque contra dicha tesis que, en uno de sus muchos rostros, apunta a un recodo raramente tanteado de la concepción de lo mental y lo físico emanada de la filosofía de orientación analítica: las propias nociones de estado psicológico y estado neurológico. Si atendemos al ejercicio efectivo de la taxonomía neuropsicológica, no podremos sino advertir –o eso opinan Bechtel y Mundale (1999: 177)– que "[…] la noción de estado cerebral es una ficción filosófica". Incluso suponiendo que la relación entre estados mentales y estados cerebrales fuera, después de todo, la de realizabilidad múltiple, nada particularmente interesante se desprendería de ello en cuanto atañe a la relación entre "actividades, funciones y mecanismos" (Wilson y Craver 2007: 99) psicológicos y sus trasuntos neurológicos, es decir, en cuanto atañe al objeto de estudio real de la neurociencia cognitiva.

Pero es más bien la impugnación del concepto de estado –cerebral o psicológico– lo que parece una huera filigrana argumental. En su sentido técnico, que –es cierto– se baraja más en la reflexión filosófica sobre el sentido y alcance de la ciencias cognitivas que en la propia investigación empírica en ese ámbito, el término "estado" –como atinadamente extracta García-Carpintero (1995: 43)– se refiere sencillamente a "[…] esas entidades que aparecen como 'factores causales' en nuestras explicaciones: entidades a las que suponemos el poder para causar y ser causadas". Así, tanto eventos, sucesos o acaecimientos como procesos o estados propiamente dichos –es decir, fenómenos temporalmente estables– contarían como estados. Las actividades, funciones y mecanismos que forman la ontología teórica de la neurociencia cognitiva difícilmente puedan ser otra cosa que ordenamientos de tales estados, salvo que sean, después de todo, ficciones filosóficas. O visto desde otro ángulo: la ficción filosófica que se impugna parece ser, ella misma, el espantapájaros de una falacia flagrante.

Más fundadas se antojan las objeciones de Bechtel y Mundale (1999) a la dicotomización de niveles explicativos que parece endémica a las posiciones funcionalistas. Tan poderoso sesgo ha sido, de hecho, diseccionado en detalle por Lycan (1987), que argumenta con vigor en pro de la idea de "la continuidad de niveles de la naturaleza" (Lycan 1987: 37), o, más exactamente, en pro de una comprensión de la realidad –de marcado acentos aristotélicos– en la que ésta aparezca como:

> […] a multiple *hierarchy* of levels of nature, each level marked by nexus of nomic generalizations and supervenient on all those levels below it on the continuum. (Lycan 1987: 38)

De lo que se trata –piensa Lycan– no es de cambiar de rumbo, sino de profundizar en las intuiciones fundacionales del funcionalismo para subsanar los múltiples errores que se han derivado de una desaforada dicotomización. El propio Lycan (1987: 136) anota, de hecho, que la estructura jerárquica que intenta esbozar ya había quedado lúcidamente descrita por Simon (1969, *cf. supra*), en uno de los textos seminales de la teorización sobre inteligencia artificial y simulación cognitiva. Aun al margen de esa pujante veta, la genealogía de la propuesta es diáfana:

> Contemporary Functionalism in the philosophy of mind began with a distinction between *role* and *occupant*. […]
> […] What matters is function, not functionary; program, not realizing-stuff; software, not hardware; role, not occupant. Thus the birth of Functionalism, and the distinction between "functional" and "structural" states or properties of an organism. (Lycan 1987: 37)

Pero cuando la concepción dicotómica de la realidad es reemplazada por una concepción jerárquica como la que Lycan auspicia, al mismo tiempo que la distinción original entre función y estructura parece hacerse ubicua, sus límites se tornan permeables:

> See Nature as hierarchically organized in this way, and the "function" / "structure" distinction *goes relative*: something is a role as opposed to an occupant, a functional state as opposed to a realizer, or vice versa, only *modulo* a designated level of nature. (Lycan 1987: 38)

Entre paréntesis: aflora entonces la cuestión última –uno de los ámbitos en los que la raigambre aristotélica de esta relectura del funcionalismo se hace presente– de cómo pensar sobre los extremos de la cadena de estructura y función. Es el polo material el que preocupa expresamente a Lycan, que se muestra escéptico respecto a que podamos hallar al cabo "[…] the desired *general* mode of purely nonfunctional characterization, the vernacular of *pure occupancy*" (Lycan 1987: 48) que nos permita describir el principio indefinido al que Aristóteles aludía en *Metafísica* (Z.3: 1037a27): "We're in search of prime matter here […] (Lycan 1987: 47). El escepticismo de Lycan, naturalmente, parece instigado por la experiencia de nuestra inmersión en la estructura de la realidad, en la que, de momento al menos, cada instante en que parecemos tocar fondo se revela después como una ilusión:

> There may be "pure occupants" or prime matter, ultimately unrealized realizers, even *necessarily* fundamental particles […] but further descent is always *epistemically* possible for us, and so we have no ordinary word for pure occupancy. "Role" / "occupant" remains a level-relative distinction; all we can mean by "pure occupant" is *stuff at a level L that realizes entities of level L+1 but is not if fact realized at any lower level*. (Lycan 1987: 48)

Cómo ligar toda esta mudadiza metafísica a la rigidez que parece apropiada a la noción de causalidad sobre la que descansan principios como el de clausura o el de herencia causal, esgrimido por Kim 1993a: 355, *supra*), es una cuestión capital cuyo abordaje conviene postergar brevemente, hasta que tratemos de calibrar la distancia existente entre la idea de *poderes causales* que, como en Kim, impera en la lectura reduccionista de la tesis de realizabilidad variable y la idea de *relaciones nómicas* que resulta predominante en las posiciones antirreduccionistas, como en los trabajos de Davidson sobre el monismo anómalo.

El caso es que ya Churchland (1986) había ensayado una respuesta a las reflexiones antirreducionistas de Pylyshyn (1984) que pasaba por admitir el carácter funcional de buena parte del vocabulario teórico de las disciplinas que solemos tomar por básicas, y ensanchar de ese modo el horizonte para movimientos de reducción teórica de diferentes calados. Como concisamente rememora Bickle (2006: §2.7):

> New levels of theory thus get inserted between those describing the structure of the lower level kinds and those of purely functional kinds: between, for example, the physiology of individual neurons and cognitive psychology. We might find a common neurofunctional property for a given type of psychological state across a wide variety of distinct brains. And if the scope of the macro-theory doesn't extend beyond that of its microfunctional counterpart, then reduction will be achieved despite vast multiple realizability at the microstructural level.

En manos de Lycan, en cambio, la idea de continuidad de niveles de la naturaleza adquiere tintes que recuerdan precisamente a las propuestas de Pylyshyn (1984: *passim*) en cuanto a la viabilidad de diferentes generalizaciones bajo diferentes vocabularios teóricos (*cf*., por ejemplo, Pylyshyn 1984: 4, 11, *infra*). Un estudiante inadvertido, pongamos por caso, bien podría atribuir a Pylyshyn una afirmación como la siguiente:

> Levels are nexus of interesting lawlike generalizations, and are individuated according to the types of generalizations involved. (Lycan 1987: 38)

Sea como sea, la propuesta de Churchland y Lycan parece haber impregnado el corazón del funcionalismo cognitivista. Tanto es así que apenas una década después, por ejemplo, Block (1997: 125) –acaso uno de los pensadores que con mayor severidad ha discutido las tesis de Lycan– da por evidente que, pese a que por comodidad tendamos a hablar de *la* encarnación fisiológica o física de una propiedad psicológica, "[…] one can also consider whether biology itself is multiply realizable in physics and chemistry". Unas páginas antes, el mismo Block (1997: 108) había anotado, en el mismo espíritu, que "[…] talk of macro and micro is relative; properties that are micro relative to one set of properties can be macro relative to another". Tal vez no sea demasiado arriesgado sugerir que consideraciones de esta

índole puedan a la larga convertirse en una disolución de la polémica entre la interpretación reduccionista y la interpretación antirreduccionista del funcionalismo, si es que no lo han hecho ya, a efectos prácticos al menos. El propio Lycan (1981: 47 *apud* Bechtel 1988: 164, *infra*), más ambicioso, apunta que bajo esta óptica incluso el fisicalismo aparece como una variedad de funcionalismo.

Pues bien, los bastidores teóricos armados por Lycan (1987) pueden rescatar, por ejemplo, las expeditivas críticas de Shapiro (2000, 2004) contra la tesis de realizabilidad múltiple de topar en los mismos obstáculos que las de Zangwill (1992, *supra*). La clave de la argumentación de Shapiro es la idea de diferencias *relevantes* entre mecanismos físicos; a su entender, los ejemplos de realizabilidad múltiple paradigmáticos del funcionalismo apuntan a mecanismos físicos diferentes, pero *irrelevantemente* diferentes. Así que el antirreduccionista –piensa Shapiro– se encuentra ante un dilema: o las diversas implementaciones físicas de un sistema difieren en sus propiedades físicas *relevantes*, o no lo hacen; si no lo hacen, esas diversas implementaciones físicas no constituyen un caso de realizabilidad múltiple, *puesto que la diferencia es irrelevante*; si, por el contrario, se da el caso de que difieran en sus propiedades físicas relevantes, entonces ya no pueden considerarse instancias del mismo tipo *funcional*, por lo que, de nuevo, no constituyen un caso de realizabilidad múltiple. La aparente perspicuidad del argumento es engañosa, puesto que cada vez que Shapiro estipula, como si tal cosa, que determinada propiedad física es o no es relevante, su *fiat* oculta la apelación a las propiedades funcionales del sistema cuya continuidad dependa de dicha propiedad física. Levantado el cartón, el supuesto dilema mengua para convertirse en lo que la doctrina funcionalista tomaría como una obviedad: o las diversas implementaciones físicas de un sistema difieren en alguna de las propiedades físicas *que determinan sus propiedades funcionales*, o no lo hacen; si no lo hacen, esas diversas implementaciones físicas no constituyen un caso de realizabilidad múltiple, pero si se da el caso de que difieran en alguna de sus propiedades físicas relevantes (es decir, aquellas que determinan sus propiedades funcionales), entonces, puesto que pueden *ex hypothesi* considerarse instancias del mismo tipo funcional pero no del mismo tipo físico, constituyen un caso rotundo de realizabilidad múltiple. Que la referencia al hecho de que las propiedades físicas relevantes del sistema son aquellas que determinan sus propiedades funcionales estaba velada salta a la vista si observamos como en la primera versión del presunto dilema se pasa tranquilamente de la diferencia en propiedades físicas relevantes a la diferencia de tipo funcional, transición cuya premisa intermedia se hace explícita –y se rechaza– en la segunda versión, desbrozada, del razonamiento. En suma, el dilema enarbolado por Shapiro (2000) no refuta la tesis de realizabilidad múltiple: sólo la niega.

Sombrío o radiante, el escenario del embate de Shapiro es el tiempo imaginario en que la neurocirugía ha aprendido a reemplazar cada una de nuestras células nerviosas por un circuito integrado que, merced a su legión de transistores, desempeña al pie de la letra la función que antes ejercía la célula. Ni siquiera tal cosa

–argumenta Shapiro (2000)– sería una muestra convincente de realizabilidad múltiple, puesto que

> […] if […] each neuron's contribution to psychological capacities is solely its transmission of an electrical signal, and if silicon chips contribute to psychological capacities in precisely the same way, then the silicon brain and the neural brain are not distinct realizations of the mind. (Shapiro 2000: 645)

La acometida de Shapiro es, al tiempo, más y menos certera de lo que parece. Su contendiente –el experimento imaginario del cerebro de silicio– es un tópico de la literatura funcionalista al menos desde Pylyshyn (1980), pero también lo es de la literatura antifuncionalista desde que Searle (1992: 69) extrajera de él la conclusión de que "[…o]ntológicamente hablando, la conducta, el rol funcional y las relaciones causales son irrelevantes para la existencia de fenómenos mentales conscientes", o acaso desde que Block (1978) comenzara a alertarnos sobre las limitaciones del funcionalismo con argumentos modales parecidos. Mientras para unos es de una evidencia completamente transparente que una cirujía de esa índole aniquilaría cualquier rastro de vida psíquica en el malhadado paciente, para otros resulta clarísimo que ésta permanecería intacta. Pero Shapiro –claro está– no necesita arbitrar entre esas intuiciones contrapuestas; le basta con plantear su razonamiento bajo la hipótesis de que la sospecha funcionalista fuese la acertada.

Sea como sea, la segunda premisa hipotética estipulada por Shapiro, más que dar razón al funcionalismo, obvia precisamente una de sus reivindicaciones capitales, a saber: la existencia de generalizaciones interesantes que afloran de la descripción computacional de los procesos psicológicos con independencia del modo en que, en un nivel inferior de abstracción, se encarnen en mecanismos físicos. Dicho de otro modo: que para construir una simulación computacional completa del funcionamiento de un sistema nervioso humano no es imprescindible –y seguramente, diría el funcionalista, ni siquiera sea aconsejable– recurrir a la simulación del funcionamiento de cada una de sus células mediante un componente discreto, un circuito integrado por cada neurona según la simplificación al uso. Así las cosas, naturalmente, la conclusión sería el fruto de una rotunda *petitio*, y el embate de Shapiro resultaría fallido.

Sin embargo, el argumento de Shapiro atina –acaso de refilón– a dar cuerpo a la tesis de Lycan (1987) de que la distinción entre lo funcional y lo estructural es relativa. La idea de que hay diferencias físicas irrelevantes de cara a la descripción del funcionamiento de un dispositivo, ya se trate de una neurona o de una multitud de componentes electrónicos ensamblados sobre una placa semiconductora, es una variedad de la idea de que la distinción entre estructura y función se desplaza a merced de nuestros intereses explicativos. Considérese, si no, si las diferencias físicas entre la neurona y el circuito integrado pueden tomarse como irrelevantes de cara a la restitución de su funcionalidad en caso de daños estructurales. Pues bien, la

movilidad de las lindes entre estructura y función pertenece, como acabamos de ver, al núcleo de las enseñanzas de Lycan (1987).

Bien es verdad que Wilson y Craver (2007), por el contrario, coinciden con Shapiro (2004) en dar un alcance más ambicioso a sus conclusiones. Sabemos por von Melchner *et al*. (2000) de la reorganización intermodal del córtex auditivo de crías de hurón (*Mustela putorius furo*) en las que se han seccionado experimentalmente, en el hemisferio izquierdo, las aferencias cocleares que pasan por el tálamo de camino al córtex auditivo primario[217]. Los desafortunados hurones aprendían entonces, mediante un paradigma de condicionamiento operante, a responder discriminativamente a determinados estímulos visuales presentados en el campo visual ipsilateral a la lesión, del que se ocupan estructuras hemisféricas intactas; naturalmente, mostraban idéntica conducta que en el campo visual contralateral, dado que no se había producido lesión alguna en las vías visuales. Lo fascinante es que el núcleo geniculado medial que había sufrido la deaferenciación iba con el paso del tiempo quedando inervado de aferencias retinianas que remitía al córtex auditivo –tal como haría, intacto, con las proyecciones auditivas. Una vez asentado el aprendizaje, esos mismos animales eran sometidos a una segunda intervención, en la que se destruían las aferencias retinianas al núcleo geniculado lateral y el núcleo lateral posterior del hemisferio dañado, de manera que la nueva ruta retiniano-talámico-cortical izquierda, que aprovecha estructuras auditivas para el procesamiento de información visual, quedaba a cargo en solitario del campo visual derecho. En esas condiciones, los hurones eran capaces de mantener la respuesta discriminativa aprendida ante estímulos visuales, remedando en gran medida el rendimiento del córtex visual en la tarea, pero empleando para ello el córtex auditivo. Pero si, por último, se infligía a los hurones una lesión en el córtex auditivo primario, la detección del estímulo en el paradigma aprendido se desplomaba hasta niveles atribuibles al azar. Tenemos, así pues, una estructura cortical, las áreas 41 y 42 de Broadmann, que desempeña en el hemisferio intacto funciones auditivas y en el hemisferio lesionado funciones visuales, o, dicho de otro modo, unas funciones visuales que se implementan en el córtex visual del hemisferio intacto y en el córtex auditivo del hemisferio dañado. Pues bien: ni siquiera esto valdría a juicio de Shapiro (2004) o Wilson y Craver (2007) como un caso *bona fide* de realizabilidad múltiple –realizabilidad múltiple, se entiende, de ciertas secuencias del procesamiento visual temprano–, a la manera en la que Putnam había intentado afianzar su tesis en los hallazgos de Flourens o Lashley (*cf. infra*). La razón es que, incluso dando por bueno que las capacidades visuales desarrolladas tras la lesión sean, bajo alguna descripción, equivalentes a las del campo visual normal, dista de estar claro que haya diferencias relevantes entre el modo en que se implementan en el córtex auditivo del hemisferio lesionado y en el córtex visual del hemisferio intacto. Al contrario –

[217] En particular, se seccionan las proyecciones desde el colículo inferior, a través del *brachium* cuadrigémino posterior, hacia el núcleo geniculado medial, lo que conlleva también la ablación del colículo superior.

sostendría Shapiro (2004: 64)–, todo parece indicar que los hurones que han sufrido la reorganización neurológica inducida por la lesión logran ver "[…] sólo en la medida en que su córtex auditivo llega a parecerse a un córtex visual".

Hasta qué punto –y en qué grado de detalle– puede afirmarse que el córtex auditivo que atraviesa la reorganización intermodal se trueca en un córtex visual es una cuestión empírica extremadamente compleja que von Melchner *et al.* (2000), cuyas preocupaciones teóricas se orientan más bien a establecer que el patrón de aferencias contribuye, a lo largo del desarrollo cortical, a dar forma a las redes corticales, abordan sólo lateralmente. Pero a falta de un criterio firme en cuanto a qué diferencias serían *relevantes* y cuáles no, la posición de Shapiro (2000, 2004) y el aval que le prestan Wilson y Craver (2007) tienen la tara de hacer la tesis de realizabilidad múltiple falsa por decreto, pues no es difícil argumentar que es irrelevante, en tan vago sentido, precisamente cualquier diferencia estructural que no desemboque en diferencias funcionales. Si bien el esquema en el que engarzar ese criterio firme está disponible en el trabajo de Lycan –como el propio Craver (2001) ha resaltado–, recurrir a él debilitaría la conclusión radical que Shapiro, Wilson y Craver quieren respaldar. Así, no procedería ya insistir en que a idénticas funciones psicológicas corresponden siempre idénticos mecanismos neurológicos, sino más bien conceder que los procesos que un nivel de descripción más abstracto resultan idénticos pueden dejar de serlo en niveles inferiores –los que se tilda de irrelevantes– aun cuando sigan siéndolo en niveles intermedios –los que Shapiro estipula relevantes.

Las mismas trabas entorpecen, en el fondo, el argumento de Bechtel y Mundale (1999) según el cual la verosimilitud de la tesis de realizabilidad múltiple es una ilusión conceptual –en realidad, un "error metodológico" (Bechtel y Mundale 1999: 105)– que proviene en buena medida del doble rasero que aplicamos al discurso psicológico y al neurológico. La protesta que atraviesa el razonamiento, en la que resuenan con claridad los planteamientos de Zangwill (1992, *supra*), es que cuando admitimos que una capacidad psicológica determinada –por ejemplo, la memoria a corto plazo– pueda venir encarnada en mecanismos físicos muy diferentes –digamos, por reiterar el ejemplo de Wilson y Craver (2007), en el cerebro humano y en del pulpo– estamos describiendo la función y la estructura que la desempeña con minuciosidad muy despareja. Si los pormenores del funcionamiento psicológico de la memoria a corto plazo que diferencian a ambas especies no impiden identificarla como la misma función, tampoco –arguyen Bechtel y Mundale– tienen por qué impedir los detalles de las estructuras neurológicas implicadas que las clasifiquemos como la misma estructura. Al sumar a este doble rasero la constatación de que los alegatos antirreduccionistas inspirados en la tesis de realizabilidad múltiple tienden a simplificar en exceso las prácticas taxonómicas de la neurociencia cognitiva, obtendremos el esqueleto de la restitución del reduccionismo ensayada por Bechtel y Mundale (1999):

> […] the initial plausibility of claims to multiple realizability rest[s] on (a) mismatching a broad-grain criterion (to show sameness of psychological states) with a fine-grained

> criterion (to differentiate brain states), and (b) a failure to attend to the purposes for which taxonomies of brain and psychological states are developed. (Bechtel y Mundale 1999: 105)

La preocupación en torno a la necesidad de contar con un croquis más fidedigno de la actividad taxonómica de las ciencias cognitivas es plenamente legítima, y enlaza con la insistencia de Lycan (1987) en que renunciemos a la perezosa dicotomización entre estructura y función a la que nos hemos acostumbrado. Sin embargo, desplegar un esquema teórico más refinado, articulándolo en torno a la relatividad de dichas nociones, no supone ninguna refutación de la tesis de realizabilidad múltiple –como se ha visto ya en relación con los argumentos de Shapiro (2000, 2004).

El argumento del doble rasero, en cambio, es sumamente endeble. Que la descripción psicológica nos proporcione modos de clasificar fenómenos mentales en tipos que no estén al alcance de la descripción neurofisiológica es precisamente lo que las convertiría en descripciones epistemológicamente autónomas. En ocasiones, dichos modos de clasificación son efectivamente –como insinúan Bechtel y Mundale (1999)– más groseros que los que pueden desplegarse en el vocabulario neurológico, pero otras veces son incomparablemente más finos: difícilmente podrá calcarse, en términos neurofisiológicos, la clase de todas las creencias de que mañana hará buen tiempo, o de todos los deseos de volver a paladear el inconfundible aroma y textura de tal o cual merienda infantil… Ni siquiera se trata, en realidad, de que un vocabulario nos otorgue mayor capacidad de abstracción que otro, sino de que cada uno nos dota de distintos criterios sobre los que articular las abstracciones que encontremos explicativamente relevantes. Esa –por así decir– inconmensurabilidad entre taxonomías originadas en criterios dispares (*cf.* Kim 1978) es precisamente uno de los puntales de las lecturas antirreduccionistas del funcionalismo –no en vano, recordemos, Fodor (1985: 16, *supra*) define el funcionalismo como "[…] la doctrina de que la taxonomía teórica del psicólogo no tiene por qué resultar 'natural' desde el punto de vista de ninguna ciencia más básica". Desde la perspectiva funcionalista, la taxonomía téorica del psicólogo es explicativamente imprescindible –es decir, no es reducible a las de ciencias más básicas– debido a que aísla patrones funcionales que podrían darse en sistemas físicos suficientemente diversos como para que no hubiera modo de aislarlos en virtud de sus características físicas. Más concisamente, la irreductibilidad de la taxonomía funcional de lo mental es efecto de su realizabilidad múltiple. Pero Bechtel y Mundale (1999), por el contrario, presentan la irreductibilidad de la taxonomía funcional de lo mental –o su caricatura– como causa de lo que a sus ojos es la ilusión de realizabilidad múltiple. Pero para convencernos de que reneguemos de las disparidades taxonómicas que impugnan, tendrían que habernos convencido *antes* de que la tesis de realizabilidad múltiple es falsa. Así que lo más que puede lograr esta rama de los razonamientos de Bechtel y Mundale (1999) es *corroborar* que la realizabilidad múltiple es ilusoria si partimos de que efectivamente lo es. No está mal, pero sí lejos de demostrar la falsedad de la tesis en litigio.

Aunque por motivos distintos, Gillett (2004) concluye, igualmente, que los ataques contra la tesis de realizabilidad múltiple de Bechtel y Mundale (1999), como los de Shapiro (2000) y cuantos inciden en estrategias parecidas, dependen de una petición de principios contra las posiciones funcionalistas, de las que toma a Fodor (1974) como adalid. La clave, a juicio de Gillett, está en que Fodor y sus críticos asumen diferentes concepciones de la relación de instanciación o realización, una concepción "dimensional" en el caso de Fodor y una concepción "plana" en los planteamientos reduccionistas, incluido el de Kim. La concepción plana de la relación de instanciación, destilada en Gillett (2004: 593), asume como condiciones para que *X* sea una instanciación de *Y* que *(i)* tanto *X* como *Y* se encuentren instanciadas en el mismo particular y que *(ii)* los poderes causales propios de *Y* sean un subconjunto de los de *X*[218]. De ese modo, queda expurgada la posibilidad de que la propiedad instanciada, *Y*, se involucre en todo comercio causal que no sea aquel en que lo hace de la mano de *X*, lo que resulta crucial para la reprobación de la idea de ciencias especiales incoada por Kim. Bajo la concepción dimensional de la relación de instanciación, ambas condiciones quedan revocadas. En virtud, entonces, de la posibilidad de que la propiedad instanciada *G* y aquella en la que se instancia, *F*, no sean propiedades del mismo particular, se abre el resquicio que permite incorporar la distinción entre las partes constituyentes de un individuo y el propio individuo a nuestra descripción de la relación de instanciación. En efecto:

> Under the dimensioned view, properties/relations F_1-F_n, of the constituents of an individual *s* can play the causal role of a property *G* of *s* without F_1-F_n contributing any of the individuating causal powers of *G*. For in cases of realization involving different individuals, the realizer properties/relations instantiated in constituents do not contribute powers to the constituents that are individuative of the realized property *G* instantiated in the constituted individual. Nonetheless, F_1-F_n, play the causal role of *G* in a wider sense, for the constituted individual has the powers individuative of *G* in virtue of the distinct powers contributed by F_1-F_n to the constituents. (Gillett 2004: 594)

Lo que resulta decisivo en la relación de instanciación de *G* en F_1-F_n es, en suma, que al menos *algunos* de los poderes causales que el individuo *s* ostenta en virtud de su posesión de la propiedad *G*, los ostenta en virtud de los que él mismo, o sus partes constituyentes, albergan en virtud de su posesión de las propiedades F_1-F_n, pero no viceversa.

La distinción entre una metafísica dimensional y una metafísica plana de la relación de instanciación provee a Gillett de herramientas para elucidar dos ejemplos muy discutidos: la discrepancia entre Putnam (1967a: 228, *supra*; 1975b), por un lado, y Block y Fodor (1972a), por otro, respecto a si el ojo de los mamíferos y el de los cefalópodos constituyen un caso de realización múltiple de una misma función, y el argumento de Shapiro (2000) según el cual la respuesta a esa pregunta debería ser negativa en el caso de un sacacorchos de acero y otro de aluminio. La clave es, en

[218] La primera condición es en realidad, tal como el propio Gillett apunta, consecuencia de la segunda.

ambas controversias, que las razones para negar que nos encontremos ante un expediente claro de realización múltiple provienen de asumir una concepción plana de la relación de instanciación[219]. Las marcadas diferencias histológicas, citológicas, genéticas, etc. que existen entre el ojo de un mamífero y el de un cefalópodo quedarían así excluidas en el razonamiento de Putnam por no ser propiedades del mismo individuo –el ojo– del que se predica la función visual, sino de sus partes constituyentes; Block y Fodor, en cambio, atendiendo justo a esas diferencias, resolverían que ambos tipos de órganos son en efecto instanciaciones físicamente dispares de una misma propiedad funcional. Entre el sacacorchos de acero y el de aluminio –un ejemplo que Shapiro aduce en el contexto de su elaboración de la idea de diferencias físicas *relevantes* (*supra*)– existen también diferencias, más evidentes si cabe, que Shapiro desdeñaría por la misma razón –no atañen a propiedades del sacacorchos, sino de su estructura molecular constituyente– pese a que él mismo argumente que la razón para descartarlas es su *irrelevancia*. Es todo menos azaroso, a la vista de esto, que Shapiro (2000: 647) se alinee con Putnam en la discusión sobre los órganos visuales de mamíferos y cefalópodos.

Aunque la ofensiva de Shapiro había quedado ya sofocada al revelarse cómo la noción de diferencias físicas relevantes, sobre la que se articula, es en realidad un remedo de la de diferencias funcionales, cabe anotar ahora que ni siquiera estamos obligados a concederle las intuiciones que motivaban su evaluación de los casos en disputa, en la medida en que dichas intuiciones dependen de que asumamos acríticamente con él una cierta concepción de la relación de instanciación. También Gillett, por lo demás, incide en la idea de que el énfasis de Shapiro en la condición de relevancia es perfectamente conciliable con la lectura antirreduccionista de la tesis de realizabilidad variable contra la que Shapiro intenta argumentar –lo que no es compatible con ella, claro, es la metafísica plana de la instanciación que Shapiro presupone:

> […D]efenders of the received account of M[ultiple] R[ealization] will accept Shapiro's important insights about the relevance criterion and its underlying idea that the only properties/relations relevant to M[ultiple] R[ealization] are those in virtue of which the pertinent individual has the individuative powers of the realized property. (Gillett 2004: 599)

Naturalmente: como el propio Shapiro señala, si el antirreduccionista no estuviera dispuesto a aceptar su criterio de relevancia se vería forzado a admitir como un caso de realización múltiple, a falta de otro criterio de exclusión, cualquier conjunto de

[219] Precisamente por ello, Gillett (2004: 591) se resiste a contar a Putnam entre quienes adoptan lo que él bautiza como la "concepción heredada de la realizabilidad múltiple" –heredada, entiéndase, de Fodor (1974). Aunque esa "concepción heredada" parece coincidir con el "[…] consenso […] en cuanto a que el reduccionismo es un error" descrito por Block (1997: 107), al que Kim (1992: 1) prefiere tildar de "conventional wisdom" –ya LePore y Loewer (1989) hablaban de "received wisdom"–, lo que resulta determinante a ojos de Gillett es que Putnam parece compartir con Kim una metafísica plana de la instanciación.

sistemas funcionalmente equivalentes cuyos elementos difiriesen entre sí en *cualquiera* de sus propiedades físicas –por ejemplo, dos sacacorchos que únicamente difiriesen en el color–, vaciando de esa manera la noción de realizabilidad múltiple. Pero es un tanto estrafalario exigir al funcionalista de convicciones antirreduccionistas que acepte ese criterio de relevancia, cuando de hecho lo incluye entre los principios fundamentales de su concepción de lo mental. Implorarle, en cambio, que asuma la concepción plana de la instanciación que Shapiro da por buena es, como Gillett (2004: 600) bien se encarga de señalar, incurrir en *petitio*.

Pero la incursión de Gillett en la metafísica de la relación de instanciación, o realización, cobija frutos más jugosos que el mero desmantelamiento de la acometida de Shapiro. La distinción entre las propiedades físicas del individuo que alberga la propiedad funcional en cuestión y las propiedades físicas de sus partes constituyentes acuña en la relación de instanciación una hendidura por la que puede fluir la irreductibilidad de la explicación psicológica, al igual que lo hace Davidson (1970, 1973a, 1974a, *cf.* también 1963 y 1993), al reiterar que un mismo evento –o dos que descubramos idénticos– es susceptible de diferentes clasificaciones, y con ello de quedar engarzado en diferentes regularidades, en virtud de a cuáles de sus propiedades atendamos, así como Burge (1986, *infra*), con su énfasis en las diferencias entre individuación, causación y composición, o Pylyshyn (1984: 11, *infra*), recalcando la necesidad de distinguir entre descripción y explicación, y entre una noción de causalidad con alcance de casos y una con alcance de tipos. A la hora de la recolección, el propio Gillett hace cuentas: una vez desarmado el dilema de Shapiro, procede anotar que el antirreduccionista, si ha de mantener que puedan ser funcionalmente idénticos dos sistemas que difieren precisamente en las propiedades físicas que determinan sus propiedades funcionales, necesita recurrir a la concepción dimensional de la instanciación. En definitiva –dice Gillett (2004: 601)– dicha concepción parece inherente a la concepción heredada de la autonomía de las ciencias especiales. Si no tanto, si podemos concluir al menos que alguna suerte de palanca o alzaprima capaz de mantener abierta la fisura, el desajuste entre distintos órdenes de propiedades, es inherente a toda tesis de autonomía que haya de resultar creíble.

También, por otra parte, parece recaer en una petición de principios, teñida de apelación *ad auctoritatem*, la insistencia de Bickle (1998) en que la tesis de realizabilidad múltiple vendría desmentida por el hecho de que la actual neurociencia cognitiva asuma como principio metodológico la homogeneidad de los mecanismos neurofisiológicos subyacentes a idénticos procesos psicológicos en distintas especies, en distintos individuos de la misma especie, y en distintos momentos de la vida de un mismo organismo. Curiosamente, Bickle (2006: §2.3) pasa a renglón seguido a argumentar que los resultados empíricos obtenidos mediante diversas técnicas de imagen cerebral, que "[…] revelan áreas comunes de alta actividad metabólica durante la ejecución de tareas psicológicas, tanto en el mismo como en distintos sujetos humanos –con una resolución espacial que llega ya a un mílimetro", corroboran la falsedad de la tesis de realizabilidad múltiple. Pero es

difícil ver cómo ningún resultado empírico podría corroborar –más que trivialmente– aquello que se asume como un "principio metodológico" en la obtención de dicho resultado. Dicho de otro modo, tal vez sean verdaderos tanto el antecedente como la consecuencia del condicional retórico con que Bickle intenta desarmar los planteamientos antirreducionistas –a saber:

> If radical multiple realizability really obtained among species in the actual world, contemporary neuroscientific experimental techniques built upon this assumption should bear little fruit. (Bickle 2006: §2.3)

La dificultad, por supuesto, reside en que valorar cuánto fruto es poco o mucho no es tarea que pueda acometerse, en casos como éste, sin valorar cuán copiosa sería la cosecha si se prescindiera de los presupuestos en cuestión[220] –o, peor aún, si es que *fuera posible* prescindir por completo de los presupuestos en cuestión. No resultaría muy creíble, desde luego, si Bickle atestiguara tener una respuesta a esas preguntas, así que su alegato queda prácticamente sin efecto.

Con parecida benevolencia recoge Bickle las protestas de Bechtel y Mundale (1999) ante la asunción de la tesis de realizabilidad múltiple, planteándolas como una tajante impugnación de la "[…] consecuencia metodológica que a veces se extrae de la premisa de la tesis de realizabilidad múltiple" según la cual:

> […] if psychological states are multiply realized across biological species, then neuroscience –the scientific study of brains– will be of little use understanding cognition. (Bickle 2006: §2.3)

Ni que decir tiene que es perfectamente legítimo, como hacen Bechtel y McCauley (1999), constatar que el itinerario habitual de la investigación neurofisiológica tiene una estación casi ineluctable en los estudios sobre cerebros no humanos, y servirse de esa constatación para impulsar una versión del fisicalismo en la que las tesis de identidad psicofísica aparecen como hipótesis heurísticas urdidas para guíar la investigación más que como conclusiones de ésta. En otro contexto, Bickle (2006: §2.5) cita también la postura de Bechtel y McCauley, pero no parece reparar en que la propia naturaleza heurística de la tesis de identidad, así planteada, pule cualquier arista en la que pudiera rozarse con las propuestas funcionalistas. En efecto, la búsqueda de nexos de identidad entre procesos psicológicos y procesos neurofisiológicos bien puede ser la táctica idónea para atestiguar tanto la existencia de dichos nexos como sus límites, dondequiera que estén, y bien pudiera desembocar

[220] Un ejemplo de cuestionamiento de los presupuestos de las técnicas de imagen cerebral que, en un ámbito diferente, también podría acabar resultando fecundo, puede encontrarse en los experimentos de Sirotin y Das (2009) –véase también Leopold (2009)– sobre la relación entre actividad neuronal y flujo de sangre, la cual podría no exhibir la linealidad que se presupone en la interpretación psicológica de las técnicas de neuroimagen funcional. La razón es que el cerebro podría aumentar la irrigación de las regiones corticales en las que *prevé* mayor actividad, siendo así que tal previsión podría, naturalmente, resultar errada.

paulatinamente, o no, en la conclusión de que sus límites se hallaban más o menos allí donde el funcionalismo venía pronosticando.

Procede en este punto remedar la defensa que el propio Lewis (1969) trabara en sus primeros careos con Putnam (1967a): si el funcionalista pretendiera de hecho arrancar de la verdad de la tesis de realizabilidad múltiple, a modo de corolario metodológico, la inutilidad de la investigación neurofisiológica, las objeciones estarían bien traídas y serían eficaces, pero ningún funcionalista "razonable" compartiría semejantes aspiraciones. De hecho, como se examinará en relación con los argumentos de Field (1978: 50, *infra*), Pylyshyn (1984: 173, *infra*), o Hatfield (1989: 262, *infra*), la consecuencia metodológica combatida por Bickle (2006) de la mano de Bechtel y Mundale (1999) ha sido explícitamente rechazada por los propios funcionalistas, a quienes se les atribuye. Signo de ello es que las conclusiones de Bechtel y Mundale en cuanto a la relación entre la investigación neurocientífica y la construcción de modelos funcionales –que hay "[…] consideraciones funcionales que se incorporan al desarrollo de la taxonomía estructural y que dicha taxonomía puede a su vez constituir una guía heurística en el desarrollo de modelos de procesamiento de información" (Bechtel y Mundale 1999: 201)– no son muy diferentes de las que darían por buenas funcionalistas "razonables", incluso algunos tan reputadamente radicales como Pylyshyn.

Del mismo modo, el modesto desenlace que Bickle otorga finalmente a su argumentación en torno a la prosperidad explicativa de la neurociencia parece encontrarse dentro de los márgenes de lo que desde una posición netamente funcionalista cabe admitir –aunque quizá por distintos motivos:

> […e]ven if one accepts the multiple realizability contention, one should be hesitant to draw strong consequences about psychology's *methodological* autonomy from it. (Bickle 2006: §2.3)

En efecto, para sustentar la autonomía explicativa de la psicología no basta con establecer la verdad de la tesis de realizabilidad múltiple. La bonanza de la investigación neurofisiológica, sin embargo, no es una de las razones de que esto sea así; es natural, dado que la autonomía explicativa de la psicología no conlleva ni por asomo la indigencia de las ciencias del cerebro.

En una breve defensa de los planteamientos de Bechtel y Mundale (1999), Couch (2004) acierta a vertebrar la controversia en torno a dos cuestiones asimétricas: el valedor de la tesis de realizabilidad múltiple debe mostrar, *(a)*, que las propiedades funcionales de varios sistemas u organismos son idénticas –es decir, pertenecen al mismo tipo, especificado con criterios funcionales– y, *(b)*, que las propiedades físicas en las que dichas propiedades funcionales se materializan en los diversos sistemas u organismos *no* son idénticas –es decir, que pertenecen a distintos tipos, especificados según criterios físicos. Aportaciones como las de Zangwill (1992, *supra*), o las esbozadas por el propio Couch (2004) respecto a las diferencia funcionales entre el ojo de los mamíferos y el de los cefalópodos, habrían puesto en tela de juicio que el

primer punto haya quedado fehacientemente establecido, mientras que los razonamientos de Bechtel y Mundale (1999) o Shapiro (2000) menoscabarían nuestra confianza en que la segunda cuestión esté resuelta a favor de la tesis de realizabilidad múltiple. De cualquiera de las dos maneras, la verosimilitud de la tesis de realizabilidad múltiple queda –entiende Couch (2004)– gravemente dañada.

Pero sería precipitado concluir que conviene al reduccionismo cualquier acumulación de objeciones de una u otra índole. Alentamos un estricto reduccionismo al disputar *(b)* los indicios de verosimilitud con que Putnam (1967a) despacha su defensa de la realizabilidad múltiple, como intentan Bechtel y Mundale (1999) o Shapiro (2000), apuntando a criterios físicos o neurológicos de categorización de estados pertenecientes a un mismo tipo funcional más afinados que los que hemos logrado desarrollar hasta la fecha –siempre y cuando, claro, ese refinamiento no cargue una apelación velada a criterios funcionales ni lastres parecidos. Lo que tendríamos en tal caso es que la categorización funcional de los estados mentales y su categorización física resultan ser coincidentes si ambas se llevan a cabo en el nivel de abstracción adecuado: eso es lo que Lewis (1969: 233, *infra*) reprochaba a Putnam que ningún fisicalista sensato preconizaría, pero tal vez, después de todo, se equivocara. En cambio, abatir sólo *(a)* la idea de que sea posible agrupar en los mismos tipos funcionales estados de un organismo o sistema que resulten ser neurofisiológicamente diversos –a la manera de Zangwill (1992)– haría resquebrajarse la tesis de realizabilidad variable y, con ella, el funcionalismo, pero no impulsaría por su propia fuerza ninguna suerte de reduccionismo; antes al contrario, el aire fresco llegaría a los predios del eliminacionismo. Como se ha argumentado ya, el cuestionamiento de la pertenencia a la misma clase funcional de las diversas instancias de un estado mental debilita tanto la tesis de identidad psicofísica como pueda debilitar la de realizabilidad múltiple, ya que la desprovee de los términos psicológicos de sus pregonados enunciados de identidad[221]. Ahora bien, si resulta a la larga que tanto unas objeciones, *(a)*, como otras, *(b)*, iban por buen camino –si, en suma, el funcionalista se equivocaba doblemente, y los estados mentales no se someten a categorización funcional pero sí a categorización física–, lo que tendremos es acaso una variedad clásica de reduccionismo, afín a la tesis de identidad psicofísica en formulaciones anteriores al auge del funcionalismo: nuestros conceptos de estados mentales denotan estados cerebrales, y la propiedad de atravesar determinado tipo de estado mental es la de atravesar determinado tipo de estado cerebral, aunque nuestros conceptos de estados mentales no determinan –connotan– esa denotación suya mediante la identificación de propiedades funcionales sino que han dado en hacerlo de algún otro modo –por la apariencia introspectiva, por ejemplo. Eso, o la improbable afirmación de que nuestros conceptos mentalistas identifican tipos de estados físicos o neurológicos en virtud de las propiedades físicas

[221] Al menos en tanto en cuanto se trate de una tesis de identidad psicofísica ligada a especificaciones funcionales de los estados mentales que la flanquean: *cf. infra*.

o neurológicas de los estados a los que se refieren, siendo así que se han desarrollado de hecho en ausencia de prácticamente toda información sobre tales propiedades.

Por lo demás, la acritud de la controversia en torno a la noción de realizabilidad múltiple seguramente no sea –en eso el diagnóstico de Wilson y Craver (2007: 81) parece certero– sino un síntoma del hecho de que "[…] el concepto de realización es siervo de dos amos". Dos amos –cabría añadir– igualmente tiránicos: en una orilla, la ambición metafísica del filósofo de la mente, que –como Kim (1989, 1992, 1993)– espera extraer del concepto de realización (instanciación, etc.; *cf.* Rabossi (1995: 36, *supra*) las claves para la comprensión de la naturaleza de lo mental, su estatus ontológico y sus relaciones con lo material –*cf.* Rabossi (1995: 17, *supra*)–; en la otra, la avidez del científico por hallar goznes conceptuales capaces de engranar sus tornadizos modelos de capacidades psicológicas concretas con mecanismos neurofisiológicos igualmente concretos. Pero dado que ambos anhelos atesoran –qué duda cabe– pareja legitimidad, no parece que el concepto de realización –quizá de forma parecida a como sucediera en su día con el de reducción interteórica, o con el de identidad– pueda eludir esta tensión, sin la que, por otra parte, sería de esperar que pronto quedara inerte.

Obras o buenas razones: caridad contra herencia causal

Hemos visto, así pues, como la coherencia misma de la idea de un materialismo ajeno al reduccionismo ha venido siendo el vértice de severos ataques, que alcanzaban tal vez su mayor virulencia en el trabajo de Kim en torno a la relación de instanciación y el principio de herencia causal. Con las palabras con las que Wilson y Craver (2007: 98) recapitulan los argumentos de Kim, éstos nos habrían puesto sobre la pista de que existen razones de orden metafísico "[…] para pensar que las formas no-reduccionistas de materialismo, incluido el funcionalismo, son híbridos inestables".

Uno de los viveros en que prosperó la labor de desacreditación del antirreduccionismo emprendida por Kim es el de su prolongada disputa con Davidson a cuenta de la solidez del monismo anómalo. Es sabido que Davidson (1970) apeló a la diversidad de modos en que podemos taxonomizar un mismo suceso, o evento, para escudar la idea de eficacia causal de lo mental, y con ella la de autonomía explicativa de la psicología, ante la imputación de incongruencia con el materialismo[222]. Lo hacía –desde luego– en un *Zeitgeist* propicio: era aún cercano el eco de los trabajos seminales de Putnam (1960, 1963a, 1967a, 1967b), Fodor (1965, 1968a, 1968b) y, en un terreno algo apartado, Sellars (1956, 1963, 1969)[223]. De acuerdo

[222] Nos encontraremos de nuevo con la táctica de Davidson al examinar los argumentos antisolipsistas de Burge (1986, *infra*) y su relevancia de cara a la autonomía explicativa que el funcionalismo pueda otorgar a la psicología, así como a la integridad de la comprensión de lo mental que nos ofrece.

[223] Que Fodor conociera o no *Empiricism and the Philosophy of Mind* (Sellars 1956) se discute en Piccinini (2004) sin que ninguna conclusión firme quede asentada; puede darse casi por seguro, en cambio, que Putnam sí lo conocía.

con sus análisis, nuestras prácticas taxonómicas en relación con los sucesos mentales están inherentemente regidas por criterios de racionalidad. Así, cuando atribuimos creencias o deseos a los demás o a nosotros mismos lo hacemos *caritativamente*, tratando de salvaguardar, como atinadamente resume Liz (1995: 218), "[…] la consistencia y la completud de la vida mental de los sujetos a los que [se los] adscribimos […], así como la armonía con sus historias pasadas y sus entornos". Los sucesos pertenecientes a cualquier clase mantienen relaciones regulares, susceptibles de quedar recogidas en forma de leyes, con otros sucesos del mismo o de otros tipos: en esas *regularidades nómicas* –piensa Davidson– consiste ni más ni menos la causalidad. Un suceso mental entabla tales relaciones en tanto que no es sino un suceso físico, y lo hace, por consiguiente, sin vulnerar el materialismo ni la clausura causal que éste reclama para sí. Pero, en tanto viene descrito y clasificado en términos psicológicos, cualquier suceso mental se ve envuelto en regularidades nómicas con otros sucesos que es imposible amasar en leyes físicas, ajenas a los cánones de racionalidad que gobiernan la taxonomía psicológica. Lo que de todo esto aflora a la superficie de la actividad científica es –si Davidson está en lo cierto– que las regularidades expresadas en términos físicos pueden alcanzar carácter de ley estricta, mientras que las expresadas en términos psicológicos –impregnadas como están de consideraciones que no se refieren a relaciones brutas de causa y efecto, sino de razón y consecuencia razonable– se ven erizadas de cláusulas *caeteris paribus*: lo que aflora es, al fin y al cabo, la fisura, que separa a las ciencias exactas de las que no lo son, a las que solemos llamar "ciencias duras" de las que –con más o menos pudor– tildamos de "blandas". Desde otro punto de vista: lo mental es físico, pero es anómalamente físico. Es decir: viene clasificado según criterios tales que las regularidades que emergen bajo dicha clasificación no pueden expresarse mediante leyes físicas estrictas; tampoco pueden las propiedades psicológicas a las que esos criterios atienden –o tal vez mejor: que esos criterios conforman– quedar reducidas a propiedades físicas.

Cierta laxitud de la concepción de la causalidad que Davidson perfila puede ayudarnos a hacer frente a la constatación de que nuestra vida mental no parece tomarse muy a pecho el principio de exclusión causal que –creemos– tensa el mundo físico. La idea de que una estricta clausura atenaza la malla causal de la naturaleza subyace al principio de herencia causal que Kim nos exhorta a velar. Los votos de austeridad que así hacemos observar a nuestra noción de causa y efecto se endurecen aun más al engarzarles un principio de exclusión: de igual modo que ningún fenómeno físico, en virtud de aquel principio de clausura, puede tener una causa que no lo sea, ningún fenómeno físico podrá tener tampoco más de una causa física suficiente –o, en términos epistemológicos, más de una explicación fisicalista completa e independiente. Pero cuando se trata de explicaciones psicológicas, no es raro que una pequeña muchedumbre de ellas parezca concurrir a la comprensión de un determinado *explanandum*, y que no pocas parezcan poder dárselas de *explanans* completo e independiente de los otros. Si esto no es una mera ilusión, si efectivamente un fenómeno psicológico puede tener varias explicaciones completas y

mutuamente independientes, una manera elegante de rendir cuentas de esa peculiaridad viene dada por la noción de causalidad de Davidson, unida a su idea de que las taxonomías que rigen las regularidades nómicas de orden psicológico están irremediablemente entreveradas de criterios normativos que apelan a la racionalidad. Si bien desde una perspectiva más afín al severo reduccionismo de Kim, en la que tales regularidades quedan excluidas de la esfera legítima de la causalidad, a esta misma intuición parece ser a la que Liz (1995) apela al preguntarse retóricamente:

> ¿Por qué nuestra vida mental escapa tantas veces al principio de exclusión? ¿No será porque gran parte de las explicaciones que mencionan fenómenos mentales no son realmente explicaciones causales, sino sólo *racionalizaciones*? (Liz 1995: 228)

De cualquier manera, el hecho de que sea la urdimbre de racionalidad en la que se teje la descripción y clasificación de los fenómenos psicológicos lo que, en el planteamiento de Davidson, los dotaría de eficacia causal nos lleva a desembocar, una vez más, en un paisaje conocido: el ancho delta de la naturaleza de la intencionalidad. Como es sumamente improbable –por muchos motivos– que los criterios de racionalidad que son constitutivos de la atribución de estados mentales den en plegarse a una lógica formal, insensible al significado de sus propios enunciados, esa urdimbre de racionalidad que apresta las redes causales de lo mental estará también, inseparablemente, entretejida de intencionalidad. Entender exactamente cuál es la relación entre el deseo y lo deseado, entre la creencia y lo creído, vuelve a revelarse como la piedra angular de nuestra comprensión de la eficacia causal de lo mental –tal como ocurrió al hilo de la polémica en torno a la noción de realizabilidad múltiple (*cf.* Liz 1995: 239, *supra*)–; sobre ella habría de erguirse, también, una concepción cumplida de en qué consiste que una creencia, o un deseo, sea razón o motivo de otro, o tal vez de una acción.

También Fred Dretske (1983, 1988, 1993) ha ensayado una ruta hacia una noción de causalidad en la que puedan germinar genuinos estados mentales, y ésta parece entrecruzarse en varios lugares con la seguida por Davidson. La piedra angular del proyecto de Dretske –como bien ha señalado Liz (1995)– es la distinción entre causas desencadenantes y causas estructurantes. Si la causa desencadenante de E es C_D –si, pongamos por caso, lo que provoca un determinado movimiento es un cierto patrón de actividad nerviosa–, su causa estructurante, C_E, es aquello en virtud de lo cual C_D causa E: lo que hace que ese patrón de actividad nerviosa y no otro cause ese movimiento y no otro. Pues bien, para dar cuenta de las causas estructurantes de nuestra conducta sería imprescindible, a juicio de Drestke, hacer referencia al contenido intencional de nuestros estados psicológicos –en particular, a su contenido intencional entendido en sentido amplio, e identificado, *grosso modo*, con la función del estado psicológico en el organismo. En la acendrada destilación de Liz (1995: 234), "[…l]as *razones* de nuestra acción y de nuestros pensamientos son sus causas estructurantes".

Sea como sea, en algún punto de su razonamiento Davidson habrá de haber salvado la brecha –que, con motivo, preocupaba a Liz (1995: 212-213, *supra*)– entre la cuestión epistemológica de qué regularidades son detectables bajo uno u otro aparato conceptual y la cuestión ontológica de qué relaciones causales se dan objetivamente en la naturaleza. Pero un meticuloso escrutinio de sus pasos muestra que, en todo caso, la ha salvado sin cruzarla: Davidson, en la estela de Hume, diluye como hemos advertido la noción de relación causal en la de regularidad nómica. En el monismo anómalo, que un suceso sea la causa de otro *consiste en que* el primero pertenezca a un tipo tal que sus miembros regularmente anteceden, según un enunciado con rango de ley, a los del tipo al que pertenece el segundo. Como mucho más concisamente ha recalcado Liz (1995: 219), "[…] no hay relaciones causales sin leyes". La epistemología de las leyes es la ontología de las relaciones causales, y viceversa: la brecha entre ambos territorios, si es que se ha salvado, no se ha cruzado; no se ha cruzado porque –en este ámbito al menos– no existe.

Quedó antedicho que el quiebro de Davidson se ha granjeado abundantes objeciones. Pero la más empedernida execración del monismo anómalo es sin duda la que en las últimas décadas ha ido desplegando Kim (1978, 1979, 1984b, 1989, 1993b, 1993c). Como ya sabemos, Kim viene pertrechado primordialmente de una intuición acerca de la naturaleza de la causalidad que parece consustancial al materialismo: la de que incluso si distintas instancias de un determinado tipo de propiedades pueden quedar encarnadas en propiedades físicas de distinto tipo en distintas ocasiones, los "poderes causales" que atesoren en cada caso tendrán que quedar contenidos en los que posea la propiedad física en la que supervienen, ya que, como sin atisbo de ironía apuntan Wilson y Craver (2007: 98-99, *supra*), se trata al fin y al cabo del "mismo bulto de materia". Ése es el principio de herencia causal al que Kim (1993a: 355, *supra*) habría llamado a desbaratar la quimera de un materialismo no-reduccionista. Pero si todo esto es así, entonces tiene que haber alguna ley psicofísica estricta que describa el lazo entre los poderes causales de sucesos descritos en términos psicológicos y los de esos sucesos descritos en términos físicos, pues no dejan de ser los mismos sucesos. Y si existen tales leyes, el monismo anómalo sería un espejismo.

La réplica de Davidson (1993) a esta vigorosa ofensiva, que pasa por acusar a Kim de obviar la distinción crucial entre casos y tipos de sucesos, ha sido admirablemente condensada por Liz (1995). Desde la perspectiva del monismo anómalo, según queda trazada en Davidson (1970, 1973a, 1974a),

> […u]n mismo evento concreto puede ejemplificar propiedades muy distintas, puede ser tipificado de muchas formas. Y la identidad entre eventos concretos sólo implicaría, en cualquier caso, que esos eventos ejemplifican exactamente las mismas propiedades, todas ellas, no que tengan que ser idénticas entre sí *algunas* de esas propiedades. (Liz 1995: 219)

La observación de Davidson hace desvanecerse la aparente incontestabilidad de las intuiciones invocadas por Kim, al sugerir que reposa sobre un equívoco. Las

relaciones causales que establezca una instancia M_i (es decir, un fenómeno del tipo psicológico M) implementada en P_i (es decir, en un fenómeno del tipo físico P) no tienen por qué restringirse a las que pertenezcan a P_i aunque M_i y P_i sean un único y el mismo fenómeno. De la intuición materialista de que lo son se deriva –sí– que M_i y P_i tienen el mismo conjunto total de propiedades –puesto que son idénticos, y los idénticos, ya lo decía Leibniz, son indiscernibles–, pero no que las propiedades en virtud de las cuales lo hemos asignado a M sean las mismas que aquellas por las que lo hemos asignado a P. Bien puede darse el caso, diría Davidson, de que hayamos asignado M_i a M a tenor de unas propiedades de índole psicológica, cargadas de consideraciones normativas sobre su imbricación en un marco de racionalidad, y P_i (que no es sino M_i) a P a tenor de unas propiedades estrictamente físicas. Encuadrar a M_i en M nos abriría entonces un horizonte de generalidades nómicas válidas de los sucesos tipos M que nos estaría vedado mientras únicamente clasificáramos a P_i en P –por mucho que M_i sea P_i.

Esa propensión a atropellar la diferencia entre enunciados referidos a sucesos o a instancias concretas de una propiedad (*i.e.*, a casos) y enunciados referidos a los tipos en que se clasifican tales fenómenos vuelve a asomar en el concepto de superveniencia en sentido fuerte auspiciado por Kim (1984a). Como recuerda Liz (1995), la propuesta de Kim (1984a: 165) es que hay superveniencia fuerte de un conjunto de propiedades, A, en otro, B, si y sólo si

> […] necesariamente, para cualquier x y cualquier propiedad A_i de A, si x tiene A_i, entonces existe una propiedad B_j en B tal que x tiene B_j y tal que, necesariamente, si cualquier otro y tiene B_j, entonces también tiene A_i. (Liz 1995: 226)

Pero incluso el más desganado esfuerzo por desentrañar la lógica de esta doble condición necesaria que coimplicaría según Kim la superveniencia fuerte muestra a las claras que, de paso, implica también la identidad de tipos. De hecho, superveniencia en el sentido fuerte definido por Kim implica una variedad particularmente rígida de identidad de tipos, al convertirla no en una cuestión empírica, como imaginaba Place (1956), ni siquiera de orden teórico, como quiso Smart (1959), sino en la verdad necesaria que los teóricos de la identidad –por regla general– se han debatido para esquivar. Considérese, si el análisis no se aparece con nitidez, que A se presenta como un conjunto de propiedades $A_1 \ldots A_n$ (al igual que B: $B_1 \ldots B_n$), pero sobre cada una de esas propiedades gravita una generalización universal "para cualquier x" (o y), siendo esas generalizaciones las que operan como términos de la implicaciones materiales "si… entonces", cualificadas ambas por el adverbio modal "necesariamente". Dicho de otra manera, si se toma el cuadrado blanco como símbolo clásico de necesidad para expresar los condicionales estrictos, en el sentido de Lewis (1912), trazados por Kim, el esqueleto lógico de la superveniencia fuerte, tal como la presenta Liz (1995) no sería otro que:

$$\Box(x)(A)(\exists B)[(Ax \rightarrow Bx) \& \Box (y)(By \rightarrow Ay)]$$

Así pues, se establece conjuntivamente que todo ente que posea la propiedad *B* (léase: una instancia de un tipo de propiedades B_i perteneciente al conjunto de tipos de propiedades *B*) posee necesariamente la propiedad *A* (*idem*), una vez antedicha la generalización universal modal recíproca (a saber: todo ente que posea la propiedad *A* posee necesariamente la propiedad *B*). Dado que *x* e *y* se deslizan sobre un mismo rango de sujetos –por lo demás, no especificado–, si desmembramos la conjunción, redistribuimos los cuantificadores, y unificamos las letras de variables, obtenemos:

$\Box(x)(A)(\exists B)(Ax \rightarrow Bx)$
$\Box(x)(A)(\exists B)\Box\ (x)(Bx \rightarrow Ax)$

El segundo de los enunciados muestra ya el alcance de tipos que acaso la forma conjuntiva velara, pues predica que para toda propiedad de tipo A existe una propiedad de tipo *B* tal que para todo *x*, si *x* posee dicha propiedad de tipo *B* posee también una de tipo *A*. No es, sin embargo, un enunciado de identidad, sino una mera implicación material. El primero, en cambio, muestra sólo alcance de casos, pues deja la cuantificación universal sobre *x* fuera del alcance de la cuantificación existencial sobre *B*, y predica sólo que para todo *x* y toda propiedad de tipo *A* existe una propiedad de tipo *B* tal que si *x* posee *A* también posee *B*. Nuevamente, se trata sólo de una implicación material, recíproca respecto de la otra. La diferencia de alcance entre ambos juicios de implicación es lo único que impide condensarlos en uno de coimplicación que, a su vez, pudiera interpretarse como signo de una relación de identidad.

Así leída, la noción de superveniencia fuerte hilvanada por Kim muestra una asimetría insólita: se trata de una concepción de las relaciones entre lo mental y lo físico según la cual para cualquier clase de fenómenos mentales hay una clase de fenómenos físicos tal que todo sujeto que atraviese un fenómeno físico de esa clase alberga también un fenómeno mental de aquella, y todo sujeto que albergue un fenómeno mental de aquella atraviesa también algún estado físico, pero no necesariamente uno de esa misma clase en todos los casos. Una vertiente de la intuición que sostiene la tesis de realizabilidad múltiple queda pues respaldada (a saber: dos estados –procesos, propiedades, etc.– mentales del mismo tipo pueden venir encarnados en dos estados –procesos, propiedades, etc.– físicos de distinto tipo), mientras que la vertiente contraria queda desmentida (a saber: no es cierto que dos estados físicos del mismo tipo puedan encarnar sendos estados mentales de tipo diferente).

Sin demérito de su posible verosimilitud, es incuestionable que dicha noción de superveniencia fuerte presupone entonces la falsedad –si bien una falsedad sólo parcial– de la tesis de realizabilidad múltiple. En consecuencia, cualquier argumento que tomara como premisa la superveniencia fuerte para establecer la falsedad empírica de las conclusiones antireduccionistas que se desprenden de la tesis de realizabilidad múltiple quedaría convertido en una flagrante *petitio*. Naturalmente,

no hay *petitio* si lo que se procura con la noción de superveniencia fuerte es demostrar una verdad necesaria por medio de una argumentación de índole conceptual –como la insinuada por Rabossi (1995, *supra*)–; es decir, si se pretende acreditar que la superveniencia fuerte es necesariamente verdadera y que, en virtud sólo de ello y el buen uso de reglas lógicas tan elementales como el principio de no contradicción, podemos deducir que la realizabilidad múltiple es necesariamente falsa. Ahora bien, dista de estar claro que el argumento de Kim pueda atesorar tal pujanza, toda vez que sus premisas distan de ser –por así decir– nociones comunes como las enumeradas por Euclides en el primer libro de los *Elementos*.

Huelga decir que si la diferencia de alcance que se ha detectado no fuese más que una ilusión –digamos– conceptual, generada por alguna mínima imperfección de la expresión en lenguaje natural de una tesis lógica, o de la interpretación de dicha expresión, las cosas podrían cambiar en uno u otro sentido. Cabe, por una parte, la posibilidad de que ambos términos de la conjunción, al eliminar ésta, debieran ser entendidos con alcance de tipos. Es decir, que fuera preciso someter el cuantificador universal que afecta al sujeto, *(x)*, al alcance del cuantificador existencial que fija la existencia de la propiedad (B_j de) *B*, *(∃B)*, como ya lo está (y), para alcanzar una cabal comprensión de espíritu de la propuesta de Kim. En tal caso, lo que obtendríamos es que necesariamente, para cualquier propiedad (A_i de) *A* existe una propiedad (B_j de) *B* tal que para cualquier *x*, si y sólo si *x* tiene (A_i de) *A*, *x* tiene (B_j de) *B*. De ese modo, la noción de implicación material, que teníamos duplicada, habría quedado reemplazada a la postre por la de equivalencia:

$\Box(A)(\exists B)(x)(AxBx)$

Pero esto sería a todas luces más fuerte, dada la cualificación modal que abre la expresión, que la propia definición de identidad de tipos que Liz (1995: 223, *supra*) condensara en la expresión lógica *(M)(∃F)(x)(MxFx)*[224]. De cualquier forma, el mero hecho de que la noción de superveniencia fuerte valiera tanto como la de identidad de tipos –ni siquiera, por matices modales, más que ella– ya la enfrentaba al dilema de la fehaciente demostración a priori o la petición de principios. Es obvio que acrecentar su fuerza modal no hace sino apuntalarla en esa problemática posición.

Por otra parte, sin embargo, cabe también que lo correcto sea dar a ambos términos de la conjunción alcance de casos, deshaciendo la asimetría en sentido

[224] No es raro, obviamente, puesto que, como se vio *supra*, la debilitación de la tesis de identidad con alcance de tipos para darle alcance de casos consiste precisamente en la operación inversa a la que acabamos de describir: verbigracia, en introducir el cuantificador existencial referido a la propiedad física con la que se identifica una propiedad mental bajo el alcance del cuantificador universal referido a los posibles sujetos de la propiedad mental –formalmente, como se dijo, anteponer el cuantificador universal al existencial. Bajo esta interpretación, entonces, que el resultado del análisis de la noción de superveniencia fuerte de Kim resulte en términos lógicos más fuerte que la formulación canónica de la tesis de identidad de tipos provendría tan sólo, a fin de cuentas, del uso que Kim hace de la noción de necesidad para matizar modalmente su postulado.

opuesto. De hecho, la definición de superveniencia fuerte de Kim permite un pequeño margen de ambigüedad, que Liz ha resuelto en favor de la lectura analizada, pero que podría también volcarse del lado contrario dando pie a una interpretación en esta línea. Lo que Kim asevera es exactamente que:

> A *strongly supervenes* on B just in case, necessarily, for each x and each property F in A, if x has F, then there is a property G in B such that x has G, and *necessarily* if any y has G, it has F. (Kim 1984a: 165)

Es posible que la diferencia con la lectura de Liz sea tan decisiva como sutil. A la vista de lo que sigue a la conjunción en la definición de Kim, la formalización de la condición de superveniencia fuerte que conviene podría no ser la que venimos manejando, sino más bien:

$\Box(x)(A)(\exists B)(Ax \rightarrow Bx) \& \Box(y)(By \rightarrow Ay)$,

donde la reiteración del operador modal es un artificio retórico. Más al pie de la letra, si se quiere, en lo concerniente al hecho de que los tipos de propiedades F y G pertenecen a su vez a géneros de orden superior, A y B:

$\Box(x)(F \ni A)(Fx \rightarrow (\exists G \ni B \& Gx)) \& \Box(y)(Gy \rightarrow Fy)$

F y *G* –tanteemos el terreno un momento para no perder pie– serían propiedades psicológicas o físicas, respectivamente, tomadas como tipos: *F* –digamos– el conjunto de instancias de propiedades psicológicas que consisten en sostener determinada creencia, y *G*, el conjunto de instancias de propiedades físicas que consisten en albergar un determinado patrón de activación en cierta región cortical. *F*, pues, pertenecería a su vez –junto con muchas otras– al género *A* de propiedades psicológicas, y *G* –por su parte, y también bien acompañada– al género *B* de propiedades físicas.

Si esta interpretación es cabal, la eliminación de la conjunción arrojaría:

$\Box(x)(A)(\exists B)(Ax \rightarrow Bx)$
$\Box(x)(A)(\exists B)(Bx \rightarrow Ax)$

Ergo:

$\Box(x)(A)(\exists B)(AxBx)$

De modo que lo que Kim estaría caracterizando como superveniencia fuerte sería en realidad una relación de identidad de casos de orden necesario, y el cargo de *petitio* estaría fuera de lugar.

También Liz (1995: 227) achaca a la noción de superveniencia fuerte de Kim una "excesiva fortaleza". Sin embargo, su cuasi-reducción de la superveniencia fuerte a la identidad de tipos procede de acuerdo a un protocolo sustancialmente distinto. El argumento de Liz no depende de ninguna lectura en particular de las ambigüedades presentes en la formulación de la tesis de Kim. Por el contrario, Liz recurre a la estipulación de propiedades físicas disyuntivas, del corte de las que Block y Fodor (1972a, *supra*) anatematizaron, como términos de enunciados de identidad de tipos[225]. Tal vez por ello, la conclusión de que "[…l]a eficacia causal de las propiedades mentales acaba siendo *reducible* (idéntica, coextensiva con) la estricta eficacia causal física de ciertas propiedades físicas [se sobrentiende: disyuntivas]" no se le antoje a Liz (1995: 227) particularmente preocupante, máxime cuando puede esgrimir, en defensa de la eficacia causal de lo mental bajo esas condiciones, la reversibilidad de la relación de identidad que tanto irritara, en cambio, a los eliminacionistas:

> […T]al vez lo que mejor asegure la eficacia causal de lo mental sea la identificación de las propiedades mentales con ciertas propiedades físicas. […N]o hay razones contundentes para negar que, entonces, lo mental dejaría de ser causalmente eficaz en cuanto mental. Pues, de darse esa identidad, por razones análogas podríamos también decir que lo físico dejaría de ser causalmente eficaz en cuanto físico. (Liz 1995: 239)

Con otras palabras: si en virtud de su identidad con lo físico, lo mental cede su eficacia causal en tanto que mental, entonces en virtud de su identidad con lo mental, lo físico deja también ha de cederla en tanto que físico. Luego o nada es causalmente eficaz *per se* –lo que, salvo que queramos regresar a Malebranche o a Leibniz, raya en la *reductio*– o hay razones para afirmar que lo mental no dejaría de ser causalmente eficaz en cuanto mental[226].

Pero este análisis formal acaso ilumine menos el fondo del asunto que la discordancia entre la jerga de las regularidades nómicas en que se hace preciso repasar los argumentos de Davidson y las de los poderes causales que prefiere Kim. La noción de *poderes causales* evoca con claridad una concepción realista de las relaciones causales, en las antípodas de la idea humeana de causalidad que trasluce en el concepto de regularidad nómica. Es de suponer que los poderes causales de un estado físico –pues es de estados de lo que, según la propuesta de Kim (1993a: 355, *supra*), se predica la posesión de poderes causales– sean propiedades suyas, es decir, presumiblemente, propiedades de la entidad física que atraviesa el estado en

[225] La cuestión que queda pendiente, claro, es si las disyunciones imaginadas serían en algún sentido ordenadas, si cabría encontrarles alguna suerte de principio rector, o, por el contrario, resultarían tan desmandadas que la anhelada reducción resultaría, al menos en términos epistemológicos, contraproducente –una cuestión que, en efecto, se ha escrutado con cierto detenimiento *supra*, al sopesar la controversia entre la concepción del funcionalismo de la que es epítome el pensamiento de Fodor y aquella, más afín al fisicalismo, de la que lo es el de Lewis.

[226] Que es en realidad lo contrario de lo que asevera Liz, aunque parece que involuntariamente, en el marasmo de una triple negación.

cuestión, o, más verosímilmente, propiedades de dichas propiedades. Poderes causales de un cuerpo sólido, por ejemplo, podrían ser su fragilidad o tenacidad, su maleabilidad, su aleabilidad, su conductividad eléctrica o térmica, su óptica, su punto de fusión o ebullición etc.: propiedades disposicionales, que describen las reacciones físicas que experimentaría el cuerpo ante distintos tipos de estímulos, o, sobre todo, los efectos que pudiera originar en otras agrupaciones de moléculas en diversas circunstancias físicas. Pero parece claro que estas propiedades son derivados de unas propiedades elementales, atómicas o subatómicas, que atañen a la estructura molecular que forma el cuerpo en cuestión. Sería poco caritativo imputar a Kim la idea de que los poderes causales de un estado físico sean propiedades de primer orden de ese estado, adicionales a sus propiedades elementales y de índole relacional –eso nos abocaría, una vez más, a las viejas cualidades ocultas que, según denunciara Newton en su *Óptica*, "[…] ponen una barrera al desarrollo de la filosofía natural" (Newton 1704: 346). No obstante, el empleo del giro "poderes causales" referido a estados físicos (o mentales) parece dar a entender que se alude a alguna suerte de propiedades actuales, que revisten o no al estado en cuestión en cada momento dado. Un ejemplo elocuente de esa forma de entender la causalidad –que bien podríamos llamar, remedando a Ryle (1949), paramecánica– trasluce en Bickle (2006: §2.7, *supra*), cuando se da por buena, como premisa en un argumento según cuya conclusión cabe esperar que exista un mecanismo molecular universal para cada clase de estados psicológicos (salvo epifenómenos), la tesis de que "[…i]n the end, any psychological kind that affects an organism behavior must engage the cell-metabolic machinery in individual neurons", puesto que "[…i]n the brain, *causally speaking, that's where the rubber meets the road*" (énfasis añadido). Así, puesto que los poderes causales son propiedades actuales de un sistema, han de tener por así decir un *locus* propio en el sistema, y el nivel de descripción en el que dicho *locus* se ubique –para Bickle, el de la bioquímica– se convierte en el explicativamente privilegiado. Además, de ese sobrentendido se desprende entonces la idea de que tales poderes causales son para cada "bulto de materia" los que son, con independencia de los criterios bajo los que taxonomicemos su estado, los efectos que en distintas circunstancias lo siguen, las causas que lo anteceden, y las propias circunstancias mencionadas.

Con qué fuerza cabe atribuir eficacia causal a propiedades funcionalmente definidas vuelve a vislumbrarse, entonces, como la cuestión crucial. En ese sentido, acierta Lycan (1987) al anotar que los estados físicos que instancian las propiedades funcionales relevantes para un tipo de estado psicológico –digamos, la creencia de que *p*– vendrán sin duda determinados por propiedades neurológicas de orden superior con relación a niveles más elementales de explicación física: la apariencia de que el establecimiento de vínculos causales es natural en el ámbito de esas propiedades neurológicas pero no lo es en el de las propiedades funcionales[227] sería,

[227] Donde Levin (2004) ve la principal ventaja de la tesis de identidad de especificaciones funcionales defendida por Smart, Lewis y Armstrong: *cf. infra*.

entonces, una ilusión provocada por la perspectiva. Asignar eficacia causal a las propiedades neurológicas, según esto, exigiría una flexibilización de la noción de causalidad tanto como pudiera requerirlo asignársela a propiedades funcionales –o como mucho, una iteración menos del mismo expediente de flexibilización.

Cuando la noción de relaciones causales no se vierte en la de poderes causales inherentes a las cosas mismas sino, como en Davidson, en regularidades nómicas que operan en el seno de una concepción humeana de la causalidad, la cosa cambia. Un mismo "bulto de materia" puede ahora participar en diversas regularidades en virtud de cuáles de sus propiedades –elementales o no– se tomen como andamiaje para clasificarlo, de cómo clasifiquemos las causas y los efectos que intervienen, también, en la regularidad en cuestión, así como las circunstancias en qué ésta se da, y, particularmente, de en qué grado exijamos que la ley que expresa la regularidad tenga un carácter estricto. Parece, en suma, que el fondo de la contienda entre Davidson y Kim es una discrepancia de orden metafísico acerca de la naturaleza de la causalidad, y que no es sino esta discrepancia lo que explica los relieves de la controversia. Venimos a dar, por esta ruta, en la conclusión opuesta –o más bien, complementaria– de la que, al hilo de la misma polémica, alcanzábamos de la mano de Liz (1995): si antes conveníamos en que las perspectivas de un análisis adecuado de la causalidad mental estribaban más en la elucidación "[…] de nuestras concepciones semánticas que de nuestras concepciones acerca de la causalidad" (Liz 1995: 239, *supra*), ahora, una vez desentrañado en parte el litigio que enfrenta a Kim y Davidson, encontramos que las distintas concepciones de la causalidad mental que uno y otro preconizan con pareja lucidez se nutren de distintas concepciones de la causalidad *simpliciter*. Se nos impone, pues, el dictamen de que ni la cuestión de la naturaleza de la intencionalidad de lo mental y su eficacia causal ni la de la naturaleza de las relaciones de causa y efecto llegarán a aclararse del todo si no es juntamente.

Conviene apuntar, además, que nuestra idea de la causalidad y nuestra idea de lo mental sólo parecen confluir cuando ambas se rinden a una severa desustancialización. No deja de resultar significativo, en este sentido, el patente parecido entre la crítica de la noción de poderes causales recién esbozada y el austero análisis disposicional al que Ryle (1949, *supra*), en el apogeo del conductismo lógico, sometiera al concepto de lo mental para depurarlo de las excrecencias cartesianas que a su juicio lo infectaban. Resulta irónico que de un ardid parecido se valiera precisamente Descartes, convencido de la realidad de la substancia incorpórea del alma, para tratar de hacer ver que el misterio que envuelve a la influencia recíproca del alma y el cuerpo no es mayor que el que reviste cualquier fuerza física, cualquier poder causal como la mismísima gravedad. El 21 de mayo de 1643, después de dejar dicho que la unión del cuerpo y el alma es, como el pensamiento, la extensión o el número, una de esas "[…] nociones primitivas, que son como originales en cuyo patrón nos basamos para construir todos nuestros demás conocimientos" (Descartes, *Correspondencia con Isabel de Bohemia y otras cartas*: 28), Descartes había intentado convencer a Isabel de que concebir cómo el alma mueve al cuerpo no es más

laborioso que concebir cómo la gravedad mueve a los cuerpos hacia el centro de la Tierra –si bien, por supuesto, hemos de usar para ello nociones distintas: "[…] la gravedad, que no es nada que pueda separarse en realidad del cuerpo" (*ibid.*) no debe ser explicada mediante la idea de *unión* con la que entendemos los efectos del alma sobre el cuerpo o el cuerpo sobre el alma. Durante el verano de 1648, el penúltimo de su vida, Descartes recibe en París dos cartas de Antoine Arnauld, el autor del cuarto conjunto de objeciones a las *Meditaciones Metafísicas*, y, en su respuesta, profundiza en la misma idea:

> La mayor parte de los filósofos, que creen que la gravedad de una piedra es una cualidad real distinta de la piedra, creen que entienden bien cómo esa cualidad puede mover la piedra hacia el centro de la tierra, porque creen tener una experiencia manifiesta de ello. En cambio, yo, que estoy convencido de que no hay una cualidad tal en la naturaleza, ni por consiguiente ninguna idea verdadera de ella en el entendimiento humano, creo que ellos se sirven de la idea que tienen de la substancia incorpórea para representarse esa gravedad; de manera que no nos resulta más difícil a nosotros entender cómo la mente mueve el cuerpo, que a ellos cómo tal gravedad impulsa la piedra hacia abajo. Y no importa que digan que esa gravedad no es una substancia, pues en realidad la conciben como una substancia, ya que creen que es una cosa real, que por medio de cierta potencia (la divina) podría existir sin la piedra. Tampoco importa que crean que es corpórea, pues si por corpóreo entendemos todo lo que atañe al cuerpo, aunque sea de otra naturaleza, entonces la mente también puede llamarse corpórea, en cuanto que es apta para unirse con el cuerpo; pero si por corpóreo entendemos lo que participa de la naturaleza del cuerpo, esa gravedad no es más corpórea que la mente humana. (Descartes, *Correspondencia con Arnauld*, 209-210, 29 de julio de 1648, en *Meditaciones metafísicas y otros textos*.)

No habría, pues, un problema de la mente y el cuerpo en mayor medida en que hubiera un problema del cuerpo y el cuerpo. El aparente misterio sería sólo fruto de la aplicación de nuestros conceptos al ámbito equivocado, del reiterado mal uso de nuestras facultades de percibir, imaginar y concebir, de nuestra desmedida ambición de *explicar*. Así pues, de tener razón Fodor en que el funcionalismo habría logrado conciliar la insistencia conductista en "[…] el carácter relacional de las propiedades mentales […]" con el énfasis fisicalista en la "[…] autonomía ontológica de los particulares mentales y, con ello, el carácter causal de las interacciones mente-cuerpo" (Fodor 1981a: 9, *supra*), habría que matizar que la admirable maniobra exige más contorsiones teóricas de las que parece a simple vista. En particular, nos exige plegarnos a una noción de causalidad que acaso haya quedado más debilitada de lo que estemos en condiciones a conceder –máxime teniendo en cuenta los múltiples propósitos que tal noción está llamada a servir en cada uno de los ámbitos de la ciencia y de la vida cotidiana. De lo contrario, el carácter causal de las propiedades psicológicas y su carácter relacional quedarían reconciliados de un modo muy incompleto: casi totalmente desligados entre sí, atado en corto el primero –como sospecha Kim– a las propiedades físicas de los estados que en cada caso particular instancien un estado mental de un determinado tipo, vinculado en cambio el

segundo –como reitera Davidson– a los criterios según los cuales el estado mental en cuestión ha sido asignado a ese tipo y no a otro.

Con todo, la argumentación de Davidson parece haber tomado otros senderos. A su juicio, si el monismo anómalo ha de escapar del acecho del epifenomenismo es adhiriéndose a un principio de superveniencia debilitado, según el cual a toda diferencia en propiedades mentales subyace una diferencia en propiedades físicas, y salvaguardando a la par una distinción nítida entre unas y otras. Este es el difícil equilibrio que Liz describe con sobriedad como la tesis de que:

> […l]as propiedades mentales serían causalmente relevantes porque tener o no tener ciertas propiedades mentales, o tener o no tener propiedades mentales en general, implicaría *cambios* en las propiedades físicas, y estos cambios en las propiedades físicas *sí* son causalmente relevantes. (Liz 1995: 221)

Resulta claro, sin embargo, que esta táctica tensa en exceso tanto el espíritu de la noción de superveniencia, según ha venido siendo articulada por Kim, como el proclamado carácter monista del planteamiento de Davidson; Liz (1995: 222) ha tomado nota diligentemente de las reiteradas objeciones que Kim (1978, 1979, 1984, 1989, 1993b, 1993c) y Sosa (1993) han hecho valer al respecto. Por una parte, la idea de que las propiedades mentales puedan implicar cambios en las propiedades físicas que a su vez alteren el orden causal del mundo puede tomarse en dos sentidos, según entendamos por *implicar* una relación conceptual, como en *entrañar*, o causal, como en *generar* o *provocar*: en el primer caso, no sería fácil sostener que hallamos dado con un nicho de eficacia causal propiamente mental –estaríamos más bien ante un paisaje ya conocido: o tanto lo mental como lo físico abrigan la anhelada eficacia causal, o ésta elude tanto a lo uno como a lo otro– ; en el segundo caso, si es que hemos hallado tal yacimiento, tendremos que convenir en que su relieve no se compadece bien –salvo tal vez muy a primera vista– con la tesis de que todo cambio en una propiedad mental *sea* un cambio en una propiedad física, o en un conjunto de ellas –pues lo que se haría ahora difícil es explicar cuál sería la causa de los cambios en las propiedades mentales en las que arraigaría la presunta eficacia causal de lo mental, si no fuese una previa alteración de las propiedades físicas en las que éstas supervienen. La tesis de superveniencia no sobrelleva bien una lectura tan forzada. Pero, además, trasladar el peso de la argumentación a la distinción entre propiedades físicas y propiedades mentales, haciendo residir en estas últimas la capacidad inexplicada de alterar el orden del mundo aunque sólo por mediación de las primeras, hace quebrarse la propuesta ontológica de Davidson del lado de un rotundo dualismo de propiedades –que no en vano acaba recordando en mucho al propugnado por Popper y Eccles (1984). En otras palabras: si Davidson acepta que la controversia se libre en el terreno de la concepción de la causalidad que Kim da por buena, parece claro que lleva las de perder.

Alguna suerte de extensión –por distensión– de la idea de causalidad parece, en definitiva, necesaria si ésta ha de acompañar a nuestra concepción de lo mental,

incluso bajo el cielo despejado de la realizabilidad múltiple. La preocupación que aflige a Liz (1995) es que la tesis de realizabilidad múltiple, de ser correcta, desmantele "la relevancia explicativa y eficacia causal de lo mental" (Liz 1995: 231). Esto sucedería, a su entender, porque la verdad de dicha tesis nos haría ver que el hecho de que según criterios psicológicos asignemos determinados fenómenos a una cierta familia es inerte en términos causales[228], y por tanto explicativos, en tanto en cuanto dichos fenómenos son físicamente dispares, y en tanto en cuanto es sólo su naturaleza física lo que condiciona su comportamiento causal, y por tanto explicativo. Para ese mal, la extensión de la idea de causalidad –"[…] *ampliar el concepto de causalidad* a fin de dar cabida a explicaciones y relaciones causales establecidas a través de propiedades irreductiblemente funcionales" (Liz 1995: 231)– es uno de los remedios que Liz considera, y rechaza. No es prudente, sin embargo, asumir ese rechazo a la vez que se da por bueno que la falsedad de la tesis de realizabilidad múltiple abriría la puerta de la eficacia causal de lo mental, y con ello de su relevancia explicativa, precisamente al dar alas a la idea de identidad entre tipos: es obvio que una eficacia causal de lo mental que fuera indistinguible de la cosechada por lo físico no comportaría relevancia explicativa alguna –por no hablar de irreductibilidad–, y es difícil siquiera ver en qué sentido podría considerarse propiamente *eficacia*, siendo su naturaleza enteramente vicaria. Así que impugnar la realizabilidad múltiple de lo mental por el expeditivo trámite de aceptar enunciados de identidad entre tipos de fenómenos psicológicos y disyunciones de tipos de fenómenos físicos –otra de las tácticas para evitar la ineficacia causal y la irrelevancia explicativa de lo mental que Liz ensaya y rechaza–ni siquiera tiene visos de acercarse al propósito esbozado, sin perjuicio de los problemas de diversa índole que suscita y que ya estudiaran Block y Fodor (1972a).

El camino que permanece franco para Liz, entonces, pasa por aceptar la tesis de realizabilidad múltiple "[…] *resistiéndose, al mismo tiempo, a abrir la puerta de la causalidad a las propiedades funcionales*" (Liz 1995: 231), lo cual sería factible si admitimos que "[…] las *propias* explicaciones y relaciones pretendidamente causales que mencionan fenómenos mentales serían *múltiplemente realizables*" (*íbid*.). Dicho al detalle: si P_1 y P_2 son instancias de propiedades psicológicas pertenecientes, según criterios psicológicos, al tipo P, pero P_1 es idéntica a la propiedad física F_1, perteneciente según criterios físicos al tipo F y P_2 es idéntica a la propiedad física F'_1, perteneciente según criterios físicos al tipo F', entonces la eficacia causal de P_1 será de distinta naturaleza de la de P_2, pues la de P_1 dependerá en realidad de F_1 y la de P_2 de F'_1; la relevancia explicativa de la asignación de P_1 y P_2 a P residirá entonces en su capacidad de –digamos– atravesar la distancia entre F y F' para alumbrar el hecho de

[228] "Inerte" es también la expresión que suele emplear Fodor (1989, *infra*); "impotente", en cambio, escriben Jackson y Pettit (1990: 203) en el contexto de una discusión acerca del estatus causal de las disposiciones (*cf. infra*), mientras Dietrich (2005: 124), más recientemente, se propone combatir la tesis de que "[…] las descripciones semánticas de estados mentales son *ociosas* desde la perspectiva de la psicología cognitiva" (énfasis añadido). Se trata, en definitiva, de si de acuerdo con el funcionalismo determinados aspectos de lo mental resultan ser epifenoménicos.

que F_1 muestra bajo cierta noción de causalidad C un comportamiento causal equiparable al que muestra F'_1 bajo cierta noción de causalidad C', puesto que el comportamiento causal de F_1 bajo C hace de P_1 una instancia de P, al igual que el comportamiento causal de F'_1 bajo C' hace de P_2 una instancia de P. Una propuesta de este género parece, a simple vista al menos, dotada de la envergadura precisa para abarcar en su seno tanto la severa concepción de la causalidad que Kim, o el propio Liz, parecen favorecer como la blandida por Davidson, mucho más maleable. Por otro lado, la conclusión reduccionista que Liz, con aire descorazonado, extrae de todo esto es que:

> […] la ilusión de que lo mental pueda ser siempre, por sí mismo, explicativamente relevante y causalmente eficaz, *con independencia de las peculiaridades físicas* de los sistemas concretos en los que se encuentre ejemplificado, se desvanece. (Liz 1995: 231)

Pero este desaliento parece a todas luces infundado. Una vez que hemos transigido con distintas concepciones de la causalidad, bajo las cuales se detectan distintas regularidades, es gratuito restringir la idea de eficacia explicativa a los casos en que una generalización permite enlazar regularidades que afloren bajo distintas ideas de causa. Por mucha que sea la relevancia que eso confiera a una explicación, no resta ninguna relevancia a las regularidades detectadas bajo una concepción de lo causal determinada, sin cruces con ninguna otra. De hecho, a la vista de argumentos como los desplegados por Davidson, no es descabellado esperar, antes bien, que las regularidades que logremos aislar bajo una concepción de la causalidad en la que tengan cabida ciertas consideraciones normativas, expresadas en leyes *caeteris paribus*, resulten de por sí particularmente relevantes, al punto de que lleguemos acaso a considerar justificado adjudicar genuina eficacia causal a las propiedades que nos hubieran permitido identificar dichas regularidades, o a los eventos en que se instancian.

Aparejos para apresar lo mental

Es atribuible a la perspicacia de Skinner (1974: 7) el esclarecimiento primero de que el conductismo –*su* conductismo, al menos– "[…] no es la ciencia de la conducta humana; es la filosofía de esa ciencia". Como sucintamente dejan anotado O'Donohue y Kitchener (1999) al hilo de la posición de Zuriff (1985):

> Behaviorism […] is not the science itself, but rather the meta-position in which basic questions about what is the proper subject matter of psychology and how this subject matter should be properly studied are raised […]. (O'Donohue y Kitchener 1999: 10)

Un paso más –de la filosofía de la ciencia a la gnoseología– nos incita a dar Ringen (1990), evocando las palabras con las que Skinner (1990) rendía tributo a Bertrand Russell –la lectura de cuyos *Fundamentos de Filosofía* (1927), que cuando llegó a

Harvard atesoraba al lado de las obras capitales de Watson y Pavlov, contribuyó decisivamente a moldear las inquietudes intelectuales de Skinner, según él mismo (Skinner 1979: 10) registraría. Al igual que uno de los propósitos cruciales de la vida de Russell habría sido desvelar los límites de lo que puede ser conocido, Skinner (1990: 103) anotaba que "[…] one of mine has been to discover what it means to be a knower". No obstante –protesta Ringen–:

> Surprisingly little discussion of the nature of this Skinnerian concern exists in the literature on Skinner's life and works. Yet, it is as a theory of knowledge that the nature and significance of radical behaviorism is most clearly exhibited. (Ringen 1990: 160)

Otro tanto –con matices, claro– se predica del cognitivismo. Además, desde luego, de una gnoseología, hay en el cognitivismo, particularmente en la medida en que su silueta queda perfilada contra la del conductismo –en la medida en que dialoga con Skinner– una cierta filosofía de la ciencia, cuya marca de fábrica, según venimos viendo, viene a ser una decidida defensa de la autonomía explicativa de la psicología. Como no podría ser de otro modo, hay divergencias y heterodoxias en este aspecto: no en vano cabe incluso que planteamientos declaradamente cognitivistas vayan de la mano de inquebrantables compromisos eliminacionistas, que coincidirían en que el orden psicológico de explicación es enteramente irreductible a otros más básicos pero lo es –subrayarían– sólo en virtud de su radical, irremediable falsedad o, peor aun, de su incoherencia; tal es el caso de Churchland (1981, 1984, 1986). Ahora bien, el carácter cismático de tales posiciones es reconocible incluso en su retórica. Parece claro que –como se ha dicho– el cognitivismo se configura frente al conductismo como una concepción alternativa de la psicología capaz de dotar a ésta de un grado mayor de autonomía respecto de otros saberes. La reflexión en torno a la naturaleza de la explicación psicológica, de hecho, sería objeto de un interés sin precedentes según fuera asentándose el uso de los modelos de explicación amparados por el cognitivismo. Así, la investigación de Fodor (1968) sobre *La explicación psicológica* sería el hito primero del auge de una disciplina, la filosofía de la psicología, cuya nieve virgen apenas había hollado Wittgenstein.

Por otra parte, es sabido que el énfasis de Skinner en el carácter epistemológico –las más de las veces, en realidad, parcamente metodológico– del conductismo disimulaba a menudo su reticencia a asumir abiertamente las repercusiones ontológicas de sus planteamientos, que quedaban, en consecuencia, tan vehementemente reivindicadas aquí como austeramente desligadas del núcleo de la posición defendida, allá. Pero el cognitivismo –particularmente, si se quiere, en la medida en que dialoga con Ryle–, es una toma de posición, en un sentido peculiar, francamente ontológica, que se presenta y trata de hacerse valer como concepción de lo mental –como una novísima concepción de lo mental, de hecho– con el mismo arrojo que como concepción de la ciencia que lo estudia. Más exactamente: la promesa fundacional del cognitivismo no sería sino la de haber dado con una concepción de lo mental *a la luz de la cual* la autonomía explicativa de la psicología

podría desligarse de todo compromiso dualista, si bien esa concepción de lo mental adopta como posicionamiento ontológico –le hemos visto ya– una acendrada neutralidad. Si, de la mano de Wundt y su ley de causalidad psíquica, la psicología se había apoyado en una u otra forma de paralelismo para emanciparse de la fisiología, el cognitivismo, tras el interludio conductista, habría venido a consagrar esa precaria independencia redimiéndola del lastre del dualismo, y lo habría logrado merced a una noción de definición funcional aprendida del conductismo lógico y de la teoría de autómatas y a la distinción entre tipos y casos aplicada a la tesis de identidad psicofísica –merced, pues, en palabras de Block (2007b, *supra*), a una metafísica de la mente. El antirreduccionismo –si todo esto es así– sería connatural a la concepción cognitivista de lo mental, o, cuando menos, constituiría su lectura epistemológica menos forzada.

La cuestión se ha planteado con frecuencia en términos de cuál sea el margen que la tesis de realizabilidad múltiple deja para la relevancia explicativa de los fenómenos psicológicos de los que se predica. Así ha quedado esbozada, al hilo de las reflexiones de Liz (1995: 212-213, *supra*), con el corolario de que no es mera relevancia, sino irreductibilidad, lo que exige la defensa de la autonomía de la psicología. Pero la tenacidad con que el cognitivismo ha custodiado ese baluarte ha despertado en algunos la sospecha de que pudiera estarse erigiendo una psicología que aspirase ni más ni menos que a desestimar los hallazgos de disciplinas que, como la fisiología, operen en niveles inferiores de abstracción –dicho de otro modo, que pretendiera desmentir el viejo lema de Johannes Müller según el cual *nemo psychologus nisi physiologus*.

En efecto, una de las preocupaciones que acaso haya llevado a no pocos teóricos afines al funcionalismo a congeniar con la interpretación de Lewis y Armstrong es la de que aceptar las tesis funcionalistas en su versión –digamos– más estrictamente funcionalista pudiera arrastrarnos hacia un altanero desdén de la investigación sobre la estructura y función cerebral, que acabara por aislar al cognitivismo de avances científicos significativos. Acaso algo así es lo que algunos temieron entender –por ejemplo– en las animosas palabras con que David C. Marr (1982) inaugurara su influyente estudio sobre la visión, articulado en torno a la identificación de los problemas computacionales que *cualquier* sistema visual ha de abordar[229]:

> […] la naturaleza de los cálculos [*computations*] que subyacen a la percepción depende más de los problemas de cálculo que deben resolverse que del soporte físico [*hardware*] particular en el que se implementan las soluciones. Expresándolo de otro modo, es

[229] No sería extraño que la reflexión de Marr estuviera inspirada en la lectura de Craik (1943: 52-53, *supra*), cuyas palabras evoca poderosamente. *Cf.*:

> It is perhaps better to start with a definite idea as to the kind of tasks a mechanism can accomplish in calculation, and the tasks it would have to accomplish in order to play a part in thought, rather than to draw analogies between the nervous system and some specific mechanism […] and leave the matter there. (Craik 1943: 52-53)

> probable que sea más fácil comprender un algoritmo si se entiende la naturaleza del problema que se está resolviendo que si se examina el mecanismo (y el soporte físico) en el que se encarna. (Marr 1982: 36)

Esa es precisamente la preocupación –avivada, como hemos visto, por las bromas de Putnam (1975b: 291, *supra*) sobre la irrelevancia del hipotético descubrimiento de que el organismo humano estuviera formado de queso suizo– que, después de respaldar la "representación abstracta de las propiedades psicológicas" como el meollo del funcionalismo, trataba de mitigar Field (1978):

> The whole point of functionalism in psychology is to provide a fairly abstract representation of psychological properties, a representation that is not tied too closely to the details of the physical structure of particular organisms; in fact, a functional theory guarantees that if two organisms are, in a suitable sense, psychologically isomorphic, then they have precisely the same psychological properties, however different they may be in those aspects of physical structure that are not relevant to establishing the psychological isomorphism. A functionalist does not say that the physical structure of an organism is *irrelevant* to its psychological properties: nearly all functionalists are materialists […]. (Field 1978: 50)

En efecto, el carácter abstracto de la explicación funcional con respecto a los mecanismos neurológicos no conlleva que la descripción de tales mecanismos sea irrelevante para la teorización psicológica. Pero el hecho sociológico mencionado por Field –que la mayoría de los funcionalistas se consideren hoy por hoy materialistas– no contribuye a ningún argumento que pueda sustentar esto. Como con más acierto plantea Hatfield (1989):

> Although commitment to an autonomous functional level of description may block reduction, it does not imply that physiology is irrelevant to psychology. […E]ven if both [algorithm] *A* and *B* could in principle be realized by quite distinct neurophysiological structures, that poses no obstacle to investigating the structure of a given visual system to see which (if either) of the algorithms it instantiates. If the neurophysiology can be seen to instantiate *A* […], that should count as evidence that *A* is the appropiate functional characterization of the […] process in that species. (Hatfield 1989: 262)

Si bien la puntualización de Hatfield es pertinente, y bienintencionada, sería erróneo ver en ella un cuestionamiento de la autonomía explicativa que el funcionalismo otorga a la psicología. Lo crucial es –desde este punto de vista– que si bien el examen neurofisiológico *puede* contribuir a establecer si es uno u otro algoritmo el que proporciona la descripción funcional adecuada de un proceso psicológico, no es una cuestión de principio que la decisión *requiera* del examen neurofisiológico –o, en una versión menos conciliadora: *es* una cuestión de principio que la decisión *no requiere* del examen neurofisiológico, es decir, que si puede tomarse fundadamente, podrá tomarse en ausencia de tal examen. La razón es, en el fondo, sencilla. Supongamos que el proceso psicológico que estamos investigando, *P*, puede identificarse con el algoritmo *A* o con el algoritmo *B*, y que ambos son funcionalmente equivalentes de

modo tal que ninguna investigación psicológica –ninguna investigación cuyos datos sean estímulos, respuestas u otros estados o procesos psicológicos– puede ayudarnos a decidir si el proceso psicológico de nuestro interés es una implementación de *A* o de *B*. Entre paréntesis: siguiendo a Hatfield, se plantea la cuestión en términos de identidad entre procesos psicológicos y algoritmos, pero, obviamente, no es difícil ajustarla a la horma al uso en los debates sobre la interpretación del funcionalismo: la identidad entre estados psicológicos –o la propiedad de abrigarlos– y estados –o propiedades– funcionales. Pues bien, ¿podría el examen neurofisiológico, en cambio, descubrir al culpable? Demos por hecho que hemos logrado detectar en los sujetos que realizan dicho proceso psicológico, cuando están realizándolo, ciertos patrones de activación, *F*, en una determinada región cortical que están ausentes cuando los sujetos se desempeñan en tareas que –según pensamos– no requieren del proceso *P*. Tan modestos como rigurosos, hipotetizamos que tales patrones de activación intervienen en *P*. Lo que necesitaríamos encontrar entonces es que hay alguna peculiaridad de dichos patrones –o bien peculiaridades anatómicas, fisiológicas, histológicas, etc. de la región que los acoge– que descarta al algoritmo *B*, y sin embargo es compatible con *A* (o viceversa). Para ello, obviamente, el algoritmo *B* habría de tener alguna propiedad funcional de la que *A* careciese, y que por alguna razón colisionara con la peculiaridad neurológica –llamémosla *N*– que hemos aislado (o viceversa). Pero eso es precisamente lo que ni *B* ni *A* pueden tener, dado que, *ex hypothesi*, *A* y *B* son funcionalmente equivalentes.

Entiéndase bien: bajo un criterio laxo de equivalencia funcional, en el que la identidad de eferencias a idénticas aferencias basta para que ésta se declare, dos algoritmos pueden ser funcionalmente equivalentes a la par que poseen propiedades sintácticas muy diferentes. Éste es –valga decir– un criterio conductista de equivalencia funcional. Pero para construir un argumento al efecto de que únicamente el examen neurológico pudiera determinar cuál de entre varios algoritmos implementa efectivamente en un organismo determinado proceso psicológico, el criterio de equivalencia funcional requerido es mucho más estricto. En particular, un argumento de esta índole sólo resulta operativo si la equivalencia funcional de *A* y *B* implica que ambos son funcionalmente indistinguibles, es decir, si no sólo comparten las mismas eferencias ante las mismas afererencias, sino también el mismo itinerario entre unas y otras. Sirva un ejemplo escolar: para convertir un valor *x* de una escala 1…8 en un porcentaje *y* podemos optar por calcular *y*=*x*/8*100, o *y*=*x*/0.08, o *y*=*x*/4*25*2, o bien por una infinidad de trámites más o menos peregrinos, no necesariamente aritméticos: todos ellos son funcionalmente equivalentes en el sentido laxo, conductista, de que arrojan indefectiblemente el mismo valor de *y* para el mismo valor de *x*, pero no lo son en un sentido más estricto que exigiera la uniformidad de –según diríamos– los pasos seguidos en el procedimiento –que, para empezar, son dos en el primer caso, uno en el segundo y tres en el tercero.

Así pues, si dos algoritmos *A* y *B* son funcionalmente equivalentes bajo el severo criterio de equivalencia exigible si el argumento ha de desacreditar la

autonomía explicativa de la psicología, entonces carecerán de desemejanzas funcionales que el examen neurológico pueda registrar; si, por el contrario, dos algoritmos *A* y *B* son funcionalmente equivalentes bajo un criterio más laxo de equivalencia, entonces poseerán desemejanzas funcionales que *tanto* el examen neurológico *como* la investigación psicológica pueden acotar. Cuestión distinta –claro está–, de orden práctico, es que sea en ocasiones una estrategia u otra la que dé mejores y más tempranos frutos.

Es de rigor señalar la pequeña prestidigitación a la Hatfield somete a la pregunta. Se plantea, en primer lugar, que varias propiedades neurológicas, N_1 o N_2, pueden instanciar la propiedad funcional *A* (o *B*) que resulte idéntica al proceso psicológico *P*, pero luego se asegura que el descubrimiento de que el sistema exhibe N_1 (y no N_2) podría permitirnos concluir que es *A* (y no *B*) la propiedad funcional que cabe identificar con *P*. Una cuestión relativa a cuál (N_1 o N_2) es la instanciación física de una propiedad funcional muta así en otra relativa a cuál es la propiedad funcional (*A* o *B*) que constituye un determinado proceso psicológico. El truco es inofensivo porque Hatfield *no* concluye que pueda darse el caso de que *sólo* el examen neurológico, que nos lleva a descubrir la presencia de N_1, pueda decidir si *A* o *B* es la caracterización funcional adecuada para *P*; su conclusión, más moderada, es sencillamente que puede darse el caso de que el examen neurológico decida la cuestión. Pero una interpretación del argumento de Hatfield que aspirase a menoscabar la autonomía explicativa de la investigación psicológica se vería obligada a salvar el equívoco.

Con independencia del argumento sopesado, conviene anotar que el criterio de equivalencia funcional instaurado por la concepción funcionalista de los procesos psicológicos es –a primera vista– sumamente estricto, dado que la especificación funcional de un proceso psicológico incluye, además de estímulos y respuestas, las relaciones con otros estados y procesos internos. Tanto si partimos para construirlo de la noción de tabla de máquina como si lo hacemos del concepto de rol causal empleado por los teóricos de la identidad psicofísica –que no en vano buscaban precisamente una forma de sobrepasar las limitaciones del conductismo lógico–, obtendremos un criterio de equivalencia muy exigente. Esto permite al funcionalista afrontar con naturalidad el escenario esbozado por Hatfield, en la seguridad de que dos procesos psicológicos implementados en algoritmos funcionalmente equivalentes en sentido estricto son sencillamente el mismo proceso psicológico (mejor dicho: son dos instancias del mismo tipo de proceso psicológico). Sin embargo, no es difícil suscitar la preocupación de que un criterio de equivalencia funcional tan escrupuloso sea, al cabo, demasiado inflexible: si para que dos (instancias de) procesos psicológicos se implementen en el mismo algoritmo, y por tanto constituyan el mismo (tipo de) proceso, es preciso que coincidan de un modo tan literal en sus patrones de relaciones funcionales con estímulos, respuestas y otros estados internos, entonces es probable que se den muy pocos tipos de procesos (o estados) psicológicos –si es que llega a darse alguno– que cuenten con más de una instancia entre sus miembros. Esto –desde luego– sería epistemológicamente

desastroso, pues abocaría al funcionalismo, también si lo entendemos como tesis de identidad de estados funcionales, al mismo síndrome de impotencia explicativa que se ha descrito respecto de su interpretación como tesis de identidad de especificaciones funcionales (*supra*).

En cualquier caso, la insistencia en la autonomía explicativa que el funcionalismo otorga a la psicología no llevó tampoco a Marr (1982) a tomar por irrelevante la investigación neurológica. Antes al contrario, hay incontables signos en su trabajo de que Marr otorgaba a la descripción de las estructuras anatómicas y los procesos fisiológicos en los que se encarna un determinado proceso cognitivo precisamente el papel al que apunta Hatfield (1989: 262, *supra*): inclinar la balanza en favor de la hipótesis de que un determinado algoritmo es el que se pone por obra cuando, en una especie determinada, se lleva a cabo tal o cual proceso cognitivo, puesto que la hipótesis rival, que explica el proceso en términos de otro algoritmo, resulta incompatible con la neurofisiología de la especie en cuestión. Ése es, por cierto, el mismo papel que mucho antes, cuando funcionalismo y cognitivismo comenzaban apenas a despuntar, le había concedido Craik (1943): la construcción de modelos mecánicos de la conducta o de los procesos psicológicos –el método sintético– debe ceñirse a las restricciones que le impone la evidencia empírica acerca de lo que biológicamente es o no es factible –es decir, a los descubrimientos del método analítico, pues de lo contrario, la simulación se convierte en mera imitación –"míseras imitaciones", diría Boring (1946: 184, *supra*).

Pues bien, el propio Marr (1982: 28) relata como años atrás él mismo había propuesto "[…] el modo en que podía implementarse en las neuronas de la retina" el algoritmo elaborado por Horn (1974) para dar cuenta de la capacidad del sistema visual de detectar qué cambios de reflectancia no se deben a fluctuaciones de la iluminación ambiental (Marr 1974); queda franca la deducción de que si cualquier hecho contrastado acerca de la estructura o función nerviosa hubiera hecho inviable la implementación del algoritmo, habríamos topado con razones de peso para reconsiderar su idoneidad. No es raro, pues, que como Gardner (1985: 325) nos recuerda en su lúcido recuento de las contribuciones de Marr, éste, después de haber logrado establecer "[…] un elegante procedimiento para computar la visión estereoscópica", propusiera reemplazar "[…] el algoritmo primitivo por otro más coherente con los datos proporcionados por la neurofisiología y la psicofísica". En efecto –asegura Marr (1982: 34)–, "[…] la elección de un algoritmo está influida por la tarea a ejecutar y por el soporte físico en el que debe actuar". Que es infundado el prejuicio de que la variedad de cognitivismo enarbolada por Marr desdeñe los hallazgos de la investigación neurofisiológica queda a la vista si reparamos, además, en la importancia que él mismo otorgaba (*cf*. Marr 1982: 42-43), a la hora de relatar el desarrollo de su manera de abordar el estudio de la visión, a resultados provenientes del ámbito de la neuropsicología clínica, como, muy especialmente, los de Warrington y Taylor (1973, 1978) acerca del distinto rendimiento en tareas de reconocimiento visual de objetos cotidianos que exhibían pacientes con lesiones

parietales en uno u otro hemisferio cerebral. No se trata, desde luego, de un recurso clandestino a la neurofisiología por parte de Marr –como se acusaba a los conductistas de escudarse en veladas descripciones mentalistas de estímulos y conductas–, puesto que la reflexión metateórica acompaña en este caso a la práctica de construcción de teorías, y lo hace con claridad meridiana:

> Si queremos comprender totalmente un sistema tan complicado como el sistema nervioso, un embrión en desarrollo, un conjunto de vías metabólicas, una botella de gas o incluso un gran programa de ordenador, debemos estar preparados para admitir diferentes clases de explicación a distintos niveles de descripción […] (Marr 1982: 29)

De igual manera que un mismo algoritmo puede tomar diversas encarnaciones físicas, un mismo problema computacional o "tarea a ejecutar" puede abordarse mediante diversos algoritmos. El razonamiento que nos aboca a esta conclusión no es distinto del que asentó originariamente la tesis de la realizabilidad múltiple de lo mental que conforma el corazón del funcionalismo –y que hemos escrutado ya, por ejemplo, en Putnam (1967a: 228, *supra*)–: "[…l]a afirmación general", apunta Marr (1982: 40), "es que debido a que diferentes animales utilizan la visión para una amplia variedad de propósitos, es inconcebible que todos los animales que gozan de ella empleen las mismas representaciones"[230]. Con eso en mente, quedan esbozados los tres niveles de análisis –físico, algorítmico, computacional– a los que Marr da primacía[231], y en cuya complementariedad insiste a menudo.

> Debe existir un nivel adicional de comprensión en el que el carácter de las tareas de procesamiento de información llevadas a cabo durante la percepción se analice y comprenda de modo independiente a los mecanismos y estructuras particulares que los implementan en nuestros cerebros. […] Este análisis no excluye una comprensión a los otros niveles –el de las neuronas o el de los programas de ordenador– sino que es un complemento necesario de ellos, puesto que en su ausencia no puede haber una comprensión real del funcionamiento de todas estas neuronas. (Marr 1982: 28)

Entre paréntesis: el esquema metateórico de Marr arroja alguna luz sobre la cuestión de si los vínculos entre el cognitivismo y la teoría de la computación son meramente circunstanciales –contingentes, como ha defendido Gardner (1985: 412, *supra*), o como Neisser (1967:6, *supra*) parecía ya insinuar al rechazar los modelos computacionales entonces al uso–, o bien algo más estrechos. Tenemos, por un lado, la constatación, como en Gardner (1985), de que el fulgurante avance de la tecnología computacional ha propiciado de hecho, históricamente, ciertos modos de pensar:

---

[230] Acaso algo menos impetuoso que Marr, o más irónico, Putnam (1967a: 228, *supra*) nos advertía de que innegablemente cabe la posibilidad de que la evolución haya conducido, en todas las formas de vida del universo, a un único correlato nervioso del dolor. "But this is certainly an ambitious hypothesis" –añadía.

[231] En la estela, como Newell (1982) o Pylyshyn (1984), de Miller, Galanter y Pribram (1960); *cf*. *infra*.

> La necesidad de comprender las tareas y las máquinas de procesamiento de información sólo ha surgido de modo muy reciente. Hasta que las personas no han empezado a soñar en estas máquinas y no han empezado a construirlas no ha habido una necesidad imperiosa de pensar en ellas seriamente. Sin embargo, una vez que los investigadores se lanzaron a especular acerca de estas tareas y estas máquinas, pronto se vio claramente que muchos aspectos del mundo que nos rodea podían beneficiarse del punto de vista del procesamiento de información. (Marr 1982: 15)

A renglón seguido, no obstante, lo que encontramos es la constatación no menos clara de que los fenómenos en los que merced a esa circunstancia histórica nos ha sido dado pensar bajo una perspectiva computacional son *en sí* fenómenos de naturaleza computacional –o, mejor: su íntegra comprensión exige la adopción, entre otras, de una perspectiva computacional–; así sería, podemos presumir, aunque la historia hubiera discurrido por otros cauces:

> La mayor parte de los fenómenos que son fundamentales para nosotros como seres humanos –los misterios de la vida y la evolución, de la percepción, el sentimiento y el pensamiento– son primordialmente fenómenos de procesamiento de información, y, si queremos llegar a comprenderlos alguna vez por completo, nuestro modo de concebirlos debe incluir esta perspectiva. (Marr 1982: 15-16)

Con todo, la salvaguarda de la autonomía de la explicación psicológica que cierta lectura del cognitivismo ha derivado de esto es en Marr titubeante. Hay en su planteamiento, qué duda cabe, una rotunda defensa de la primacía epistémica del nivel computacional de análisis, fundada en motivos pragmáticos. Así, si bien Marr considera posible "[…] al menos en principio" engarzar los distintos niveles de explicación "[…] en una totalidad integrada", lo cierto es que resulta a su juicio "[…] poco práctico enlazar los niveles en completo detalle" (Marr 1982: 29): parece que sería legítimo inferir que en caso de que culmináramos esa tarea los niveles explicativos superiores –algorítmico, computacional– se tornarían prescindibles, aunque siguiera siendo aconsejable recurrir a ellos por cuestiones prácticas. En un espíritu semejante hemos escuchado ya de boca de Marr (1982: 36, *supra*) la advertencia de que examinar físicamente el sistema que implementa un determinado algoritmo probablemente no sea tan fructífero, si hemos de entender el algoritmo, como analizar a fondo el problema al que se enfrenta.

Entender la percepción –o acaso cualquier proceso psicológico– sería entonces más difícil –o quizá sólo probablemente más difícil– si atendiéramos en exclusiva a la neurofisiología relevante, pero se diría que *en principio* prescindir de otros niveles explicativos sería factible, incluso si resultara probablemente más complicado de lo necesario. Nos hallaríamos, en suma, ante una transacción entre órdenes de parsimonia: deberíamos optar entre una explicación completa de un fenómeno que se construye recurriendo a varios niveles explicativos –y en consecuencia a una mayor diversidad de constructos teóricos–, y una explicación igualmente completa que se construye en un único nivel explicativo –*ergo*, de forma más parsimoniosa–,

pero siendo así que esta última resulta más complicada que la primera, menoscabando así su propia ventaja en términos de economía explicativa. La cuestión merecería, desde luego, un análisis más detenido, pero en este contexto hemos de conformarnos con anotar que de lo que se trataría es –por lo que se ve– de una decisión pragmática. Al mismo ámbito de consideraciones prácticas parece remitir la cautela que Marr (1982: 35) recomienda

> […] al emplear los resultados neurofisiológicos para realizar inferencias acerca de los algoritmos y las representaciones que se están empleando, especialmente hasta que no tengamos una idea clara acerca de la información que ha de ser representada y los procesos que son necesarios para implementarla.

Sin embargo, se diría que late en Marr, en otros momentos, el conato de una defensa de la idea de que los distintos niveles de explicación son autónomos por razones de principio, que desbordan esas consideraciones de orden práctico. Es entonces más que nunca cuando, en palabras de Gardner (1985: 326), Marr se perfila como "[…] un cognitivista cabal". En la manifestación más rotunda de la defensa de esa tesis se alude expresamente a la *laxitud* del engarce entre los niveles explicativos:

> Cada uno de los tres niveles de descripción tendrá su lugar en la comprensión final del procesamiento de la información perceptiva y, por supuesto, todos ellos están relacionados de un modo lógico y causal. Pero es preciso advertir un aspecto importante: como están relacionados sólo de un modo laxo, algunos fenómenos únicamente podrán ser explicados a uno o dos niveles. (Marr 1982: 34)

Es meridiano que aquí no se plantea ya una cuestión de que ciertos niveles de explicación sean más convenientes que otros por motivos prácticos, sino de que ciertos niveles de explicación sean irrenuciables debido a la naturaleza de su conexión causal y lógica con los otros niveles. Sin duda es significativo que esta toma de posición venga acompañada –si no en el esbozo metateórico, sí al menos en el relato de las vicisitudes históricas de la investigación sobre visión que han conducido a ella– del recurso a la distinción entre describir y explicar que sería capital en la articulación de la defensa de la autonomía de la explicación psicológica que debemos a Pylyshyn (1984, *infra*): en efecto, Marr (1982: 25) apunta cómo comenzó a hacerse patente, a comienzos de la década de 1970, que "[…] la tarea de [la neurofisiología y la psicofísica] […] consiste en *describir* la conducta de las células o de los sujetos, pero no en *explicarla*". Algo faltaba, pues, en el proyecto de una teoría integral del sistema visual –tal como se entendía antes de que se le incorporasen las intuiciones alentadas por la construcción de máquinas de cálculo, y, en especial, por la visión de que cabría construir una máquina dotada de visión–, y lo que faltaba vendría a ser la comprensión de que mientras una descripción de un fenómeno complejo puede ser verdadera y completa en un único nivel descriptivo, una explicación de tal fenómeno sólo puede serlo si integra los varios niveles explicativos que la complejidad del fenómeno exija.

Además de en el trabajo de Marr, interrumpido por su tempranísima muerte, buena parte del cauce del que ha manado la ortodoxia cognitivista en lo que atañe a la naturaleza de la explicación psicológica y sus vínculos con otras formas de explicación científica se halla en Pylyshyn (1984), donde el eco de Fodor (1968) es nítido. El punto de partida de Pylyshyn es la supeditación del concepto de explicación al de generalización, y ambos al de vocabulario explicativo –si bien, aunque Pylyshyn se expresa en ocasiones como si "vocabulario" estuviera en sus argumentos por *léxico*, parece estar más bien por *aparato conceptual*. Una explicación, entonces, es un intento "[…] de capturar generalizaciones, y diferentes vocabularios revelan diferentes generalizaciones"; en ese sentido "[…] las explicaciones son relativas a vocabularios particulares" (Pylyshyn 1984: 2). Es de hecho el vocabulario que emplean –en particular, ciertas propiedades conceptuales de ese vocabulario–, lo que delimita a unas ciencias respecto de otras:

> What distinguishes different sciences is that they seek to answer different questions and explain different phenomena. Frequently, however, different sciences address the *same* events, though in highly different ways. […] Often, the most distinct characteristic of explanations provided by two disciplines is the vocabulary used to express both their phenomena and their explanations. (Pylyshyn 1984: 1)

En parte quizás a tientas, Pylyshyn está cuestionando una determinada concepción ingenua de la explicación en la cual el objeto de ésta –el *explanandum*– está naturalmente constituido de antemano, y las lindes entre esos objetos naturalmente constituidos determinan la demarcación entre disciplinas. Antes bien, son los conceptos empleados en la explicación –el *explanans*– los que, de acuerdo con el razonamiento de Pylyshyn, dictan el objeto de estudio a la vez que delimitan la disciplina.

Esa concepción ingenua de la explicación puede asemejarse a la que impulsa no pocos argumentos del positivismo lógico –Hempel (1935), por ejemplo, da por incontrovertible que el objeto de una ciencia es aquello de lo que tratan sus enunciados, sin atender al hecho, apuntado por Pylyshyn, de que los enunciados de distintas ciencias bien pueden tratar de los mismos objetos utilizando para ello recursos conceptuales, o vocabularios, que alumbren distintas facetas de esos objetos hasta constituirlos de diferente modo. Las críticas esgrimidas por Pylyshyn contra esta pretendida ingenuidad, en la misma línea, pueden enlazarse con el cuestionamiento de la prioridad e independencia del vocabulario observacional respecto del teórico que cobra forma en las críticas de Hanson (1959) al positivismo lógico. En efecto, resulta claro que la mera posibilidad de fijar el objeto de una ciencia al modo positivista descansa sobre la disponibilidad de un vocabulario observacional inmaculado: serían los enunciados formulados en ese vocabulario aquellos cuya referencia coincidiría con el objeto de la disciplina en la que se integran. Con otras palabras: lo que hay de común en Hanson y Pylyshyn es que ambos rechazan que el objeto de una ciencia se identifique con la referencia de un segmento privilegiado de

su vocabulario, un segmento puramente extensional y cuyo contenido quedara fijado por la propia naturaleza del objeto en cuestión[232]. Así pues, el cotejo de los argumentos de Pylyshyn con los Hanson (contra Hempel) arroja dos frutos: uno, hace patente el modo en que la construcción de los fundamentos epistemológicos del cognitivismo se enmarca en la crisis del positivismo lógico seguramente en un grado mayor que aquel en el que el conductismo como filosofía de la ciencia se enmarcaba en el propio positivismo lógico (*cf. supra*); dos, revela en el trasfondo de la controversia una disputa acerca de si la determinación del objeto de una disciplina científica debe proceder a partir únicamente de la *extensión* de sus enunciados, o dar cabida también a su *intensión*.

Naturalmente, según el propio Pylyshyn, el caso de las ciencias cognitivas ilustraría a la perfección su tesis. Así, son los recursos teóricos propios de la explicación cognitiva los que, siempre de manera provisional, puntean los confines de la realidad cognitiva, decretando a la par el alcance del modelo explicativo en cuestión:

> Cognitive psychology is fundamentally tied to a certain class of terms which in part define the phenomena it seeks to explain (but only in part, for a consequence of explanation is that it frequently redefines its explananda) and in part dictate the sort of accounts that qualify as putative explanations. (Pylyshyn 1984: 2)

El carácter tentativo de la delimitación del *explanandum* a la luz de determinados rasgos del *explanans* hace posible que haya altas y bajas en el inventario de fenómenos llamados a someterse a un modelo explicativo dado según dicho modelo se vaya desplegando:

> As the theory develops, we find new phenomena, not originally anticipated to be part of the domain, fall naturally within the theory scheme, whereas others, originally expected to fall within the paradigm, are outside the theory scheme. (Pylyshyn 1984: 271)

Por otro lado, si la teoría ha de proveernos de conocimiento nuevo, difícilmente podremos prever, antes de su desarrollo, el trazado de las lindes que abrirá en el *explanandum*. O bien, invirtiendo la imagen del carnicero experto, que a la hora de despiezar el animal sabe hundir el cuchillo justo en la membrana sinovial, y al que Platón, por boca de Sócrates, asemeja al buen dialéctico (*cf. Fedro* 265e, *supra*): si ya supiéramos dónde están las articulaciones de la naturaleza, no precisaríamos investigar para revelarlas.

---

[232] El cuestionamiento de la prioridad natural del objeto de estudio ha dado lugar a una fructífera revisión del modo de abordar la historia de la ciencia, convirtiendo la labor en la de investigar los procesos de construcción del objeto tanto como los de construcción del discurso acerca de tal objeto. En el caso de la psicología, se trata de una aproximación que suele ligarse con los trabajos de Kurt Danziger (*cf*. por ejemplo, Danziger 1990, 1997; también Brock, Louw y van Hoorn, eds. 2004, una selección de ensayos inspirados en los principios historiográficos de Danziger).

> There is no way to determine the phenomena that will cluster together to form the object of a special science expect by observing the success, or lack of it, of attempts to build theories in certain domains. (Pylyshyn 1984: 263)

De regreso al terreno de la cognición, todo esto ha de servir, a juicio de Pylyshyn, para deshacernos de cualquier objeción al cognitivismo que se limite a señalar que éste deja sin explicar determinado rango de fenómenos, puesto que no procede exigir a la explicación cognitiva que haga justicia a nuestras opiniones ingenuas acerca de lo que debe o no formar parte de su *explanandum*. En efecto:

> Cognition, as understood and conceived as a natural scientific domain, may well not include many of the areas that pretheoretical expectations had included. (Pylyshyn 1984: 197)

Es posible, en particular, que la consciencia sea una propiedad transversal a fenómenos cognitivos y otros no cognitivos[233] y, por tanto, que esté fuera de lugar anhelar una explicación cognitiva de la consciencia:

> If even the modest success in building small-scale theories in contemporary "information-processing psychology" is any indication, we have no right to the a priori assumption that the set of conscious contents is a natural domain. […]
>
> It is by no mean obvious that, to develop explanatory theories, we must group deliberate reasoning and problem-solving processes with processes that seem to occur in a flash and with no apparent awareness that any mental activity is occurring, while at the same time leaving out such vivid mental contents as the experiences of pain, seeing a certain color, being dizzy, or feeling the pieces of a puzzle fall into place. (Pylyshyn 1984: 265)

Pero sean cuales sean los contornos de la realidad cognitiva que hayan de irse paulatinamente cartografiando, la autonomía de la explicación psicológica es una conclusión que Pylyshyn considera atinado anticipar. Sabemos que el empleo de un determinado vocabulario teórico, con los recursos conceptuales que lleve aparejados, trae consigo una cierta taxonomía de los fenómenos que conforman el *explanandum* de la teoría:

> Different descriptive vocabularies entail different structures in the space of possible events from which some particular, observed instance is assumed to be a sample. A theory never explains an entirely unique event, only an event viewed against a background of distinctions and equivalences defined by the vocabulary with which the events are described. This is what I mean when I say that theories address phenomena as "events under descriptions". (Pylyshyn 1984: 16-17)

[233] Dicho a la inversa, no cabe descartar que la la clase de procesos cognitivos sea transversal a la distinción entre procesos conscientes e inconscientes. Tampoco ha lugar, en tal caso, a los reparos que ofrece Malcolm (1971, *supra*) al cognitivismo: la posibilidad de que existan procesos cognitivos inconscientes análogos, en otros aspectos, a los que ocasionalmente se muestran a la introspección no puede ser repudiada de antemano por motivos meramente lógicos.

En este ámbito, Pylyshyn está plegándose casi al pie de de la letra al análisis de Fodor (1974: 101), prefigurado como vimos en Fodor (1968: 179):

> Every science implies a taxonomy of the events in its universe of discourse. In particular, every science employs a descriptive vocabulary of theoretical and observational predicates such that events fall under the laws of the science by virtue of satisfying those predicates.

Un científico –o, más verosímilmente, una comunidad científica– se va inclinando por cierto vocabulario teórico y eso determina su taxonomía de los sucesos que pretende explicar, o bien –Pylyshyn no se preocupa en diferenciar ambos posibles procesos– una comunidad científica elige una cierta taxonomía de lo real y esa elección cobra forma en el vocabulario teórico que tiende a preferir:

> The particular way in which one chooses to divide a set of alternatives is revealed by the vocabulary used to describe the phenomena. (Pylyshyn 1984:16)

Lo decisivo, con independencia de cuál sea la dirección del proceso y de cuáles resulten ser las fronteras exteriores de la taxonomía en cuestión, es que ésta permite la detección y explotación de generalizaciones que quedarían veladas bajo otros vocabularios teóricos –al igual que, por supuesto, oculta otras que serían visibles bajo el prisma de taxonomías o vocabularios diferentes.

Es cierto, ahora bien, que puede hacerse difícil entender la idea de que haya enunciados verdaderos acerca de un fenómeno que quepa formular en un determinado vocabulario teórico pero estén vedados en otros. Develar la razón de esa dificultad acaso la aminore: si asimilar la relatividad de la explicación al vocabulario teórico resulta arduo, sobre todo en el seno de cierta tradición epistemológica, es porque se da cierta contumacia en confundir la explicación con la descripción verídica. Pero la explicación y la descripción son empresas diferentes. Distintas descripciones del mismo fenómeno, con recurso a distintos vocabularios teóricos, bien pueden ser todas verídicas –esto es poco menos que un tópico del sentido común. Pero lo que a veces nos elude es que, siendo todas ellas verídicas, bien puede ser que una constituya una *explicación* del fenómeno y otras no:

> […] one account of a sequence of events might qualify as an explanation while another *true* account of the same sequence does not. (Pylyshyn 1984:4)

Proporcionar una explicación del fenómeno es algo más que describirlo verídicamente: es entretejerlo en una urdimbre de generalizaciones. Nada tiene de raro, pues, que la descripción de tal o cual fenómeno en un determinado vocabulario lo inserte en una malla que dé respuesta a nuestras inquietudes en torno al fenómeno –es decir, que lo explique–, mientras que otra descripción igualmente verídica del mismo fenómeno, articulada en otro vocabulario, fracase en ese empeño. Dicho de

otro modo –al hilo de una observación que Pylyshyn (1984: 4) atribuye a un trabajo entonces inédito de Block y Fodor (1972b)–, lo que estamos apuntando es que el giro "La ocurrencia de *X* (dado un cierto conjunto de generalizaciones, Γ) *explica* la ocurrencia de *Y*" genera un contexto opaco, donde la sustitución ya de *X* ya de *Y* por expresiones correferenciales suyas puede no preservar el valor de verdad de la oración, al impedir que su referente quede descrito en el preciso sentido que lo incrusta en Γ. La mera descripción de la secuencia *X*, *Y*, en cambio, se despliega en contextos transparentes, donde la ley de Leibniz no se ve vulnerada[234].

Una vez que la distinción crucial entre descripción verídica y explicación queda bien sembrada, la certidumbre de que la explicación psicológica es autónoma no es sino su fruto maduro:

> […] the entire issue of the nonequivalence of neurophysiological, behavioral, cognitive, and perhaps other modes of description rests on the fact that this discussion is conducted in the context of *explanation*, where we are concerned not just with the veracity […] but with the ability of different descriptions to *capture generalizations*, and hence with the adequacy of *explanations* couched in these different vocabularies. (Pylyshyn 1984: 4)

La autonomía de la explicación psicológica, así pues, proviene del hecho de que los recursos conceptuales que ésta proporciona permiten una caracterización del *explanandum* a la luz de la cual se hace patente un cierto entramado de generalizaciones; esas mismas generalizaciones resultan esquivas cuando el *explanandum* viene descrito –o, quizá mejor, construido– con los recursos de la explicación neurofisiológica o conductual, por verídica que pueda ser cada una de esas descripciones. Si todo esto es así, claro, no habríamos acabado por reprobar el pupilaje del conocimiento psicológico al fisiológico que defendía Müller (*supra*), pero sí por reclamar un pupilaje recíproco, de la fisiología a la psicología –confluyendo así, por cierto, en el tempranísimo alegato que al respecto avanzara William James en 1877:

> Ideas and their associations in the mind, cells and their linking fibres in the brain –such are the elements. But, whereas we directly see their process of combination in the mind, we only guess in the brain what it may be from fancied analogies with the mental phenomena. [...] In a word, if it be true that, as Johannes Müller used to say, *nemo psychologus nisi physiologus*, it is doubly true that, so far as the nerve-centres go, *nemo physiologus nisi psychologus*. (James 1877: 336)

[234] Aunque Pylyshyn (1984: 4) parece considerar esto como la *razón* de que una descripción verídica de un fenómeno pueda explicarlo y otra no, parece de rigor señalar que nos encontrarnos más bien ante una reformulación de precisamente ese hecho en términos de la concepción fregeana del significado, extendida de su ámbito originario –los enunciados en los que se predica una relación de identidad– al de la predicación de la relación *constituir una explicación de*. Naturalmente, según las propias tesis de Pylyshyn, al describir el fenómeno en el marco de esa estructura teórica lo incrustamos en una cierta red de generalizaciones, es decir, lo explicamos; es decir, a fin de cuentas, damos *razón* de él: todo en orden.

Las razones de James –a la vista está– son otras –muy otras, si se quiere–, pero lo que sustenta el argumento de Pylyshyn contra el reduccionismo fisicalista no sería entonces, a fin de cuentas, sino la imposibilidad de hacer corresponder a cada tipo de estado mental un tipo de estado físico o cerebral –es decir, la realizabilidad múltiple de los estados mentales. En efecto, la tesis de Pylyshyn –tal como la rememora Bickle (2006: §1.4, quien retrotrae su formulación hasta Fodor (1975)– es que, *dada la realizabilidad múltiple de los estados mentales*, el fisicalismo de tipos "[…] vulnera un principio elemental de metodología científica": el que nos impele a capturar cuantas generalizaciones sea posible capturar[235]. Sucede, en definitiva, que:

> […] taken under a cognitive description certain kinds of behavior enter into a system of generalizations. (Pylyshyn 1984: 11)

De esa capacidad de alumbrar generalizaciones que de otro modo nos eludirían proviene el estatus epistemológico de la explicación psicológica, que no es ignorancia provisional –*cf*. Braddon-Mitchell y Jackson 1996: 262, *supra*– ni aguja de bitácora para marearla:

> There are many reasons for maintaining that explanations of behavior must involve cognitive terms in a way that does not serve merely as a heuristic or as something we do while waiting for the neurophysiological theories to progress. One of the principal reasons involves showing that there are regularities and generalizations which can be captured using cognitive terms that could not be captured in descriptions using behavioral or physical (neurophysiological) terms. (Pylyshyn 1984:6-7)

Lo que queda por afianzar, pues, so pena de que la autonomía de la explicación psicológica quede a fin de cuentas establecida por *fiat*, es un bosquejo al menos de las razones por las que tal modo de explicación nos provee de generalizaciones que a otras se les escapan. Estas razones, según ha quedado apuntado, tienen que ver con ciertos recursos conceptuales que la explicación psicológica proporciona, con ciertas peculiaridades del vocabulario teórico en el que se articula, pero sin duda hace falta una noticia más precisa de dichos recursos y peculiaridades. Pues bien: el núcleo de la capacidad de revelar generalizaciones que sustenta la autonomía de la explicación psicológica es que su vocabulario teórico permite la descripción de sucesos tal como éstos son interpretados –percibidos, conceptualizados– por el sujeto, con independencia de cómo sean en términos objetivos. La posibilidad de caracterizar un estímulo, un estado interno del propio sujeto o una conducta suya con acuerdo al modo en que éstos se inscriben en una subjetividad, vedada en el nivel de descripción fijado por el lenguaje teórico de la neurofisiología o de la psicología de

[235] *Mutatis mutandis*, lo mismo podría afirmarse en cuanto al conductismo lógico y la multiplicidad de conjuntos de disposiciones conductuales que pudieran venir asociados a un mismo tipo de estado mental.

convicciones conductistas, es la fuente de donde mana la irreductibilidad de la explicación psicológica. En palabras de Plyshyn:

> The behavioral story […] is not equivalent to the cognitive one –for an important reason. In the behavioral story past experiences must be classified in terms of a particular, *objective* taxonomy, a taxonomy that partitions classes of histories according to the physical properties of stimulus and behaviors. The way histories must be partitioned in order for them to correspond to states of knowledge, however, requires that we be capable of speaking of such things of the meaning of a sentence, or the interpretation the person placed on a certain stimulus, or the action intended by a certain behavior that formed part of the history. Unless the history leading to a particular functional state is categorized this way, we cannot use it to explain why someone does what he does. (Pylyshyn 1984: 8)

Es, así pues, el carácter semántico del vocabulario teórico adoptado por la psicología cognitiva –significado, interpretación, intención…–, es decir, el hecho de que se halle impregnado de alusiones a la naturaleza intencional de los estados internos que postula, lo que lo desliga del rango de generalizaciones abarcables en la jerga de la física de los estímulos y las respuestas o la fisiología de la actividad neural que enlaza aquéllos y éstas, desprovistas ambas de toda huella de lo intencional. Que rendir cuentas de la naturaleza de la intencionalidad sea problemático para el cognitivismo o el funcionalismo no es, entonces, asunto menor, a la vista de que es la intencionalidad lo que sustenta el estatus epistemológico que se reclama para la explicación psicológica.

Conviene recalcar, en cualquier caso, que la autonomía reivindicada para la explicación psicológica no entraña en modo alguno que una determinada cadena de acontecimientos –una configuración estimular que afecta a las terminaciones sensoriales, una intrincada retahíla de respuestas nerviosas, un patrón de tensiones y distensiones musculares que ocasionan ciertos movimientos– no pueda ser descrita verazmente con los recursos de la fisiología y la física. Es más: las múltiples relaciones de causa a efecto que forman los eslabones de esa cadena particular bien podrían quedar descritos de forma verídica y exhaustiva en esos términos. Lo que se discute es otra cosa: que tal descripción sirva para hacer patente el hecho de que *en general* las configuraciones estimulares que son objeto de una interpretación pareja por parte de un sujeto dan lugar a movimientos equiparables desde la perspectiva del sujeto –aunque quizá muy diferentes en términos físicos–, y de que intervienen en los distintos casos acogidos a esa generalización estados internos del organismo que –por distintos que sean en lo que concierne a su fisiología– resultan semejantes en lo que concierne a su semántica[236]. Lo que se discute, en fin, es que tal descripción,

[236] El propio Pylyshyn, por cierto, anota la afinidad entre esa distinción entre conducta bajo una descripción conductual o bajo una cognitiva y la que –en el marco de las investigaciones sobre la traducción que lo condujeron al desarrollo de la *tagmémica*– acuñara en su momento Kenneth Pike (1967):

por verídica y exhaustiva que sea, constituya una explicación. Dicho de otro modo: que la descripción de ese patrón concreto de relaciones causales, entre esas causas específicas y esos efectos específicos, sea tanto como la explicación del tipo de relaciones causales que se den entre causas de ese tipo y efectos de ese tipo. En lo que atañe a la descripción de la conducta, por ejemplo:

> […e]ven though the [behavioral] description may correspond exactly to what actually happens on a particular occasion, it does not correspond to the equivalence class of all behaviors that can be systematically related to certain previous experiences. (Pylyshyn 1984: 9)

Se trata, al cabo, de una cuestión de alcance del enunciado que predica la relación causal, y lo que se defiende es que para determinadas relaciones causales –las que enlazan sucesos según la interpretación de un sujeto– dicho enunciado sólo puede adquirir alcance de tipos si se formula en un vocabulario teórico que, como el de la psicología cognitiva, acomode la noción de intencionalidad, aun cuando toda relación causal particular, todo caso concreto, pueda quedar descrito por medio de un enunciado fisicalista. Igual que hemos sabido distinguir entre la afirmación de que una instancia concreta de estado mental –este o aquel deseo, digamos, que este o aquel sujeto alberga en este o aquel instante– sea un estado neurológico y la afirmación, mucho más ambiciosa, de que todo estado mental que pertenezca a la misma clase psicológica que aquel –todo deseo, si no queremos acotar más– sea un estado neurológico perteneciente a una misma clase neurológica en la que, además, no tengan cabida estados mentales de ninguna otra clase psicológica, hemos también de diferenciar entre la afirmación de que un encadenamiento particular de causa y efecto en que ambos sean estados mentales resulte ser un suceso neurológico que pueda quedar adecuadamente descrito en el vocabulario de la neurología, y la afirmación, también mucho más ambiciosa, de que todo encadenamiento causal de estados mentales que podamos apresar bajo una generalización psicológica –esto es, formulada en el lenguaje teórico de la psicología– haya de quedar igualmente apresado bajo una generalización neurológica o, a las últimas, formulable en lenguaje fisicalista. Más concisamente: que toda relación causal concreta sea un fenómeno físico no entraña que toda clase de relaciones causales lo sea, en el sentido de que pueda delimitarse empleando el lenguaje de la física[237]. Aunque no termine de

---

> That is, the regularities that exist are to be found among perceived (cognitively described) properties or what Pike (1967) calls *emic* properties, not among objective, physically described) or *etic* properties. (Pylyshyn 1984: 13, *cf.* Pylyshyn 1984: 150, *infra*)

El trabajo de Pike se enclava en una fragosa encrucijada en el itinerario histórico del cognitivismo: dirigido por Edward Sapir, cuya concepción de las relaciones entre el lenguaje y el pensamiento habría de combatir Fodor y después Pinker, constituía un alejamiento significativo respecto de la aproximación de Chomsky a la lingüística, orientándola hacia análisis de índole más molar, si bien abriría diversos horizontes en el ámbito de la antropología cognitiva.

[237] No hacemos, se diría, una distinción adicional que requiera grados mayores de reificación de la relación causal como una tercera entidad, junto a sus *relata*; entendemos más bien, sencillamente, que

expresarse nítidamente en estos términos[238], parece claro que éste es el núcleo de su propuesta si atendemos al razonamiento que Pylyshyn, explotando el ejemplo ya tratado *supra* del peatón que contempla un accidente y llama al teléfono de emergencias, presenta como un modo "[…] más revelador de plantear la discusión sobre las limitaciones de la explicación fisiológica":

> […] the difficulty with such explanations is that they provide a distinctly different causal account linking each of a potentially unlimited number of ways of learning the emergency phone number, each of a potentially unlimited number of ways of coming to know that a state of emergency exists, to each of a potentially unlimited number of sequences of muscular movements that correspond to dialing 911. In so doing, the neurophysiological story misses the most important psychological generalization involved: regardless of how a person learns the emergency number, regardless of how he or she comes to perceive the situation as an emergency, and regardless of how the person's limbs are moved in dialing the number, a single, general principle is implicit in the entire set of these sequence[s]. (Pylyshyn 1984: 11)

Nada raro, pues, en el seno del funcionalismo: exactamente lo mismo que se afirma respecto de la caracterización de ciertos estados internos del organismo –que el carácter físico de cada uno de ellos es compatible con que las clases en que se agregan no vengan formadas según criterios físicos– se afirma respecto de las relaciones causales en las que participan dichos estados. En realidad, a duras penas podría ser de otro modo, dado que los criterios en virtud de los cuales se agrupan en clases esos estados internos tienen que ver, según el funcionalismo, precisamente con las relaciones causales en las que se inmiscuyan.

La posición de Pylyshyn es netamente funcionalista, y uno de los rasgos que lo acreditan es su esfuerzo por asegurar una tranquila convivencia entre la autonomía de la explicación psicológica y la verdad del materialismo en lo que respecta al inventario último de lo real. Con un estilo de razonamiento abiertamente inspirado en el de Putnam y Fodor, Pylyshyn deja claro que:

---

lo que se predica de los *relata* en cuanto a su taxonomización se sigue para la propia relación que los une.

[238] Pylyshyn parece reservar la distinción entre instancias y clases para los *relata*, quizá por precaución ante el marasmo metafísico al que pudiera arrastrarlo la hipóstasis de la relación, quizá porque adivine en la aplicación a la taxonomización de la relación de lo que se aplica a la taxonomización de los *relata* una redundancia de aire tautológico. Sin embargo, no tiene reparos en aplicar la distinción entre instancias y clases a las categorías que sustentan generalizaciones en las ciencias especiales y las categorías físicas, entendiendo cada categoría de una ciencia especial (que no es sino una clase de estados o eventos físicos delimitada bajo el aparato conceptual propio de esa ciencia) como una clase de categorías físicas –en particular, de acuerdo con el diagnóstico de Pylyshyn (1984: 21), una clase que no puede describirse adecuadamente con una disyunción finita de instancias de categorías físicas, es decir, una clase cuya delimitación requiere el vocabulario de la ciencia especial en cuestión. Es difícil encontrar razones por las que aplicar la distinción entre instancias y clases a relaciones debiera ser más problemático que aplicársela, recursivamente, a clases.

> Indeed, every state of a cognitive system is simultaneously a biological state and a cognitive state, since it is an element of both a cognitive- and a biological-equivalence class. (Pylyshyn 1984: 143)

En efecto, la tesis planteada por Pylyshyn, acaso en un tono más conciliador, es la misma que lleva a Fodor (1985: 16, *supra*) a plantear que el funcionalismo es "[…] la doctrina de que la taxonomía teórica del psicólogo no tiene por qué parecer 'natural' desde el punto de vista de ninguna ciencia más básica". De hecho, ambas vertientes de la tesis en cuestión –la positiva, proclamada por Pylyshyn, y la negativa, enfatizada por Fodor– encuentran su horma común en unas palabras de Fodor que el propio Pylyshyn (1984: 18) cita *in extenso*:

> If science is to be unified, then all such taxonomies must apply to the same things. If physics is to be basic science, then each of these things had better be a physical thing. But it is not further required that the taxonomies which the special sciences employ must themselves reduce to the taxonomy of physics. It is not required, and it probably is not true. (Fodor 1975: 25)

También en la estela de Fodor, Pylyshyn da por buena la idea de que una razón de que la explicación psicológica logre hacer patentes generalizaciones que son refractarias a una formulación en vocabulario fisicalista –pareja, si se quiere, al carácter semántico del vocabulario mentalista– reside en el carácter no proyectable de algunas de las propiedades a las que responden selectivamente los organismos cuya actividad se trata de explicar, donde la noción de propiedad proyectable confeccionada por Goodman (1953, *supra*) –recordemos: propiedades cuya verificación en un objeto justifica en alguna medida la expectativa de que otros objetos de la misma clase mostrarán también la propiedad en cuestión– ha quedado reescrita de manera que una propiedad es proyectable si y sólo si es susceptible de figurar en una ley física[239]. Por otro lado, el carácter no proyectable de las propiedades que intervienen en las generalizaciones propias de la explicación

[239] En un uso más lato, como el que le da Bechtel (1984: 320), no es de propiedades de lo que se predica la proyectabilidad o su ausencia, sino de clases. La tesis básica del funcionalismo aparece entonces como la de que hay clases de sucesos articuladas según criterios de orden psicológico que no son proyectables en clases de sucesos articuladas según criterios de orden físico, y la proyectabilidad se perfila como un tipo de relación entre dos conjuntos, que se verifica cuando para cada miembro de uno cabe fijar una relación biunívoca con un miembro del otro. Desde luego, si la noción de propiedad queda reformulada en términos de la noción de clase, el requisito de pertinencia respecto a una ley física mencionado por Pylyshyn viene dado: la propiedad psicológica P será proyectable –podrá figurar en una ley física– en tanto en cuanto la clase de los elementos que poseen la propiedad P sea proyectable en –sometible a relación biunívoca con– la clase de los elementos que poseen una propiedad física F, que lo es precisamente en tanto es susceptible de aparecer en leyes de la física. Bajo ciertos supuestos, pues, ambas nociones de proyectabilidad son equivalentes; estos supuestos –la traducibilidad entre las nociones de clase y propiedad, la pertinencia respecto a un determinado conjunto de leyes como criterio de adjudicación de un concepto a un determinado vocabulario teórico– son escasamente discutidos en el contexto de la controversia en la que en 1984 se hallan inmersos tanto Pylyshyn como Bechtel.

psicológica es a los ojos de Pylyshyn una consecuencia natural del carácter inferencial –interpretativo, si se quiere– de la relación cognitiva de un organismo con su entorno. De esta manera, el hecho de que esa relación entre el sujeto y su mundo diste de la omnisciencia –es decir, el mero hecho de que sea posible el error– asoma una vez más como el fundamento de cuanto la explicación psicológica tiene de singular. Así:

> Another way of putting the matter is to say that organisms can respond selectively to properties of the environment that are not specifiable physically, such properties as being beautiful, being a sentence of English, or being a chair o a shoe. These properties are not properties involved in physical laws, they are not *projectable properties*. […I]t is not surprising that an organism reacts to nonphysical or nonprojectable properties of the environment inasmuch as "reacting to the environment" […] typically involves such processes as drawing inferences to the best available hypothesis about what, in fact, is out there in the distal world. (Pylyshyn 1984: 15)

Una razón –se ha dicho– de la autonomía de la explicación psicológica es el carácter no proyectable de ciertas propiedades del entorno a la que los organismos responden selectivamente. No se trata, sin embargo, de la única razón: Pylyshyn (1984: 152) deja claro que, contra la opinión de Fodor (1984), considera la no proyectabilidad como un criterio suficiente para determinar que la explicación de la actividad de un organismo debe abordarse bajo parámetros psicológicos, pero no como un criterio necesario. Incluso en los casos en que un organismo responde a propiedades de su entorno que resultan proyectables sobre propiedades físicas, cabe aun que esa respuesta venga regida por representaciones e inferencias en otros puntos de su desarrollo, haciendo imprescindible el vocabulario teórico de la psicología para rendir cuentas de ella. De hecho, eso es lo que sucede, según Pylyshyn (1984: 165), cuando consideramos la actividad del organismo desde las fases tempranas del procesamiento perceptivo, pues sobre el funcionamiento de los transductores pesa un requisito de proyectabilidad de las propiedades a las que son sensibles[240].

Es precisamente al referirse al concepto de transducción cuando Pylyshyn se muestra más explícito respecto al papel que concede a la investigación neurofisiológica en la construcción de explicaciones psicológicas –si bien sus conclusiones parecen claramente extensibles a cualquier otro ámbito de las relaciones entre neurociencia y psicología cognitiva, fuera del estudio del procesamiento perceptivo temprano. Al socorrido consuelo de la función heurística se añade la tesis de que la evidencia neurofisiológica, como vimos en Marr (1982) y se adivinaba ya en Craik (1943, *supra*), puede vetar posibles explicaciones psicológicas por depender de mecanismos incompatibles con ella. Esto, por supuesto, queda lejos de decretar irrelevante la investigación fisiológica, pero también de considerar sus resultados

[240] Habrá ocasión de analizar ese requisito en cierto detalle más adelante, al abordar la cuestión de cuál sea el vocabulario idóneo para caracterizar estímulos y respuestas en la teorización psicológica.

suficientes, o siquiera necesarios, para afianzar desarrollos teóricos en el ámbito propio de la psicología.

> This does not mean that biological data of various kinds are irrelevant to decisions about the transducers an organism has. Although it is neither necessary nor sufficient that we discover neural loci for particular transducers (no particular empirical data by themselves are ever necessary and sufficient to establish a theoretical construct), such evidence is often useful, especially when it is considered along with psychophysical and behavioral data. In addition, general neurophysiological considerations often can be used to narrow the set of possible mechanisms likely to be satisfactory. (Pylyshyn 1984: 173)

Así que, igual que ocurría en Marr (1982, *supra*), no es tan fiera la autonomía explicativa de la psicología como a veces se ha temido que fuera. Cabe, en suma, dar por probado que una concepción funcionalista de lo psicológico no entraña –no al menos con la fuerza de la necesidad lógica– menosprecio alguno de la capacidad de hacer avanzar nuestra comprensión de la mente que la fisiología tiene ampliamente acreditada, ni siquiera si dicha concepción se interpreta bajo parámetros antireduccionistas. El desarrollo de esa taxonomía teórica en el seno de un nivel de explicación intermedio "[…] entre la psicología ordinaria de las creencias y los deseos, por un lado, y la explicaciones neurológicas […], por otro" (Fodor 1985: 15, *supra*) –que, como se ha reiterado de la mano de Fodor (1985: 16, *supra*), no está forzada a coincidir con la de las ciencias más básicas– y la plena autonomía de esa "representación abstracta de las propiedades psicológicas" (Field 1978: 50, *supra*) carecen de consecuencias en cuanto atañe a la viabilidad del estudio fisiológico de la misma realidad, así como en cuanto a la fecundidad de una y otra perspectiva a la hora de fertilizarse recíprocamente.

Es de rigor tomar nota, antes de atender a otras cuestiones, de una puntualización respecto a lo que los planteamientos de Pylyshyn entrañan de cara al alcance explicativo del vocabulario conductista. La equiparación que Pylyshyn da por buena entre el vocabulario en que el teórico conductista pretende describir estímulos y respuestas –aunque luego, en la práctica, se valga de descripciones mentalistas encubiertas: Pylyshyn comparte, como se ha visto, las protestas de Chomsky al respecto[241]– y el vocabulario en que vendrían descritos en una física del estímulo o en una cinemática de la conducta no hace justicia a la pluralidad de concepciones de la explicación psicológica que conviven en la región que hemos dado en llamar "conductismo": no, cuando menos, al modo en que el propio Skinner aconsejaba caracterizar estímulos y respuestas. Como ha quedado ya explicado, Skinner –cuya postura difícilmente podremos desestimar por poco representativa del conductismo, habiendo sido, de todas todas, su más enardecido defensor– se inclinó desde muy al principio de su carrera por la definición funcional tanto de estímulos como de respuestas, es decir, por su clasificación en tipos de acuerdo con criterios

[241] *Cf.* por ejemplo Pylyshyn (1984: 8-9, *supra*). También Miller, Galanter y Pribram (1960: 233, *supra*) denuncian las tretas del conductismo en este sentido.

funcionales, renunciando así en la práctica a su caracterización fisicalista. Así, por ejemplo, cuentan para Skinner como estímulos de cierto tipo todos aquellos que acrediten cierto efecto regular sobre la conducta del organismo, con independencia de sus propiedades físicas. Éste es, de hecho –como también se ha argumentado *supra*–, el origen de las vigorosas objeciones blandidas por Chomsky (1959) contra la vacuidad de algunos de los ensayos de explicación de la conducta verbal por parte de Skinner (1957).

Sea como sea, la cercanía entre las seminales críticas de Chomsky (1959) al programa de investigación de Skinner (1957) y las desplegadas por Pylyshyn es incontestable. Así ha quedado ya de manifiesto al analizar el modo en que Pylyshyn contrarresta la posibilidad de que el conductista reasimile la interpretación cognitivista del condicionamiento articulada por Brewer (1974) y la reintegre en un esquema explicativo de estímulo–respuesta: Pylyshyn concede que el conductista puede hacer tal cosa, pero sólo merced a que los conceptos de estímulo y respuesta hayan quedado definidos bajo criterios de radical inespecificidad, y ése es un argumento nítidamente chomskiano. No menos marcada –cabe añadir– resulta la huella de Chomsky en las objeciones de Pylyshyn al realismo perceptivo directo de Gibson:

> If that is what is directly picked up [people, furniture, animals, the property of being far away, being honest, "affording" warmth or nourishment], there is, trivially, no problem involved in explaining perception. Indeed, there is no problem involved in *psychology* in general, since we directly pick up the causes of our behavior. In direct realism, it seems, the cognitivist's displaced homunculus has at least found a home in the information pick-up function. This –what Fodor and Pylyshyn (1981) call the "trivialization problem"– is the single most serious problem a direct realist confronts. (Pylyshyn 1984: 182-183) [242]

La trivialización del problema de la percepción en el realismo perceptivo directo –en Gibson como en Reid, en Reid como en Ockham, en Ockham como en casi toda la tradición aristotélica, pese a Alhacén– es la misma falta en que incurre el conductismo radical al desplazar inadvertidamente la carga explicativa de unos procesos psicológicos bajo impugnación a unos estímulos y respuestas sometidos a una violenta, aunque cuidadosa, torsión conceptual, en virtud de la cual recogen cuanto de epistemológicamente problemático pudieran tener los procesos psicológicos impugnados, pero salvaguardando, por un vehemente *fiat*, su inmaculado carácter fisicalista. Ésa es exactamente la falta que Chomsky reprochara a

[242] La crítica exhaustiva de los planteamientos de Gibson a la que Pylyshyn se remite se halla en Fodor y Pylyshyn (1981), y su conclusión bien puede cifrarse, como hace el propio Pylyshyn (1984: 270), en la idea de que "[…] having a percept is not the sort of state a system conceivably can achieve by "resonating" to some property of the environment". Esta formulación, por otro lado, hace patente el vínculo con la lectura que hace Pylyshyn (1984: 13, *supra*) del trabajo de Hochberg (1968), con la caracterización del cognitivismo como cierta suerte de pesimismo –frente al optimismo que compartirían conductistas y gestaltistas– y, por esa vía, con la cuestión de la pobreza del estímulo que se ha intentado aquilatar en relación con las propuestas de Malcolm (1971: 387, *supra*).

Skinner. La sobrecarga mentalista de los conceptos skinnerianos de estímulo, respuesta o reforzador es la sobrecarga mentalista de los objetos naturales del entorno gibsoniano. En palabras de Pylyshyn:

> In a behavioral analysis, recognition of a stimulus, response, and reinforcer is presupposed in ways that are outside the theory; that is, a behavioral analysis cannot give an account (based on conditioning) of what constitutes a stimulus, response, or reinforcer. Thus these are objects the organisms must individuate (or pick out or encode) prior to subsequent conditioning. Thus, in behavioral terms, the organism's environment consists of precisely those entities. (Pylyshyn 1984:180)

En la medida en que –como se ha defendido *supra* de la mano de Boring (1950: 667), Marx y Hillix (1963/1979: 195), MacKenzie (1977), Yela (1980/1996: 172), Gondra (1992: 15-16) u O'Donohue y Kitchener (1999: xx)– el conductismo sea una entidad de naturaleza proteica, bien cabría acusar genéricamente *al conductismo* tanto de (en algunos de sus avatares) una excesiva rigidez fisicalista en la definición de estímulos y respuestas, que proscribe explicaciones fructíferas de orden distinto, como de (en otros) una excesiva labilidad en dicha definición, que remeda la potencia de esas explicaciones afectando fidelidad a una epistemología fisicalista. Dado lo proteico de la propia personalidad de Skinner como teórico, no es descabellado incluso acusar a Skinner mismo de ambas negligencias. Ahora bien: si la aparente iniquidad de tal juicio ha de quedar diluida, sólo puede ser haciendo patente que el imputado no es presa simultáneamente de ambos errores, sino que oscila entre uno y otro. Más ecuánime, vistas así las cosas, sería reconocer la propuesta skinneriana de caracterización funcional de estímulos y respuestas como un avance significativo hacia la concepción cognitivista de la explicación psicológica, avance que se vio lastrado por la obstinación de Skinner en no incluir en sus teorías referencias a estados o procesos internos –una interpretación, por otro lado, que acaso soliviantara más que a nadie al propio Skinner.

En realidad, la renuencia de Skinner a mencionar estados o procesos internos en la teoría psicológica tiene que ver con el hecho de que en el conductismo radical se da una actitud inconmensurablemente más severa que en el cognitivismo por cuanto hace al papel de la fisiología en la explicación de la conducta, unida a un talante epistemológico, de raigambre baconiana (*cf.* Ringen 1990, *supra*), más proclive a la descripción que a la explicación o a la formulación de teorías. En un desmedido embate –que, de rendirnos a su literalidad, nos obligaría a prescindir de la práctica totalidad de las explicaciones que hemos cosechado en cualquier ámbito de la ciencia–, Skinner (1950) cargaba contra:

> […] any explanation of an observed fact which appeals to events taking place somewhere else, at some other level of observation, described in different terms, and measured, if at all, in different dimensions. (Skinner 1950: 69)

Eso censuraba, por supuesto, toda explicación de la conducta de un organismo en términos de sus estados o procesos mentales, pero también en términos de un "Sistema Nervioso Conceptual" (Skinner 1950: 70) –claramente afín a las propuestas cognitivistas–, o bien de procesos fisiológicos o electroquímicos. Si no, antes, de la disputa entre Watson y Loeb (*supra*), el cognitivismo heredaría de Skinner –como él de Crozier (*supra*)– la convicción de que la conducta de los organismos puede quedar explicada en un vocabulario independiente del articulado por la fisiología, aunque fuera modificando sustancialmente la caracterización de dicho vocabulario. En realidad, tras un escrutinio más cuidadoso, bien parece que los pioneros del cognitivismo habrían *aprendido* de Skinner a sustentar la autonomía explicativa de la psicología valiéndose de la noción de relaciones funcionales. Pero la convicción de que hay un nivel autónomo de explicación –conductual, psicológico– quedaría en el cognitivismo despojada de las desproporcionadas directrices de aislamiento entre disciplinas con que Skinner había guarnecido su filosofía de la ciencia –el conductismo radical– sin que su ciencia –el análisis experimental de la conducta– le conminara a ello. De ese modo, los cognitivistas se permitirían hablar con igual desenvoltura del carácter antireduccionista y del carácter interdisciplinar de su programa de investigación. La gestación de una nueva criatura académica y administrativa, las "ciencias cognitivas" –a imagen de las "ciencias de la vida" que ya habían prosperado en más de un campus, *cf*. Martel Johnson (1997: 3), pero también de las "ciencias de la conducta" tal como las entendía, por ejemplo, James G. Miller (1955, *supra*)– sería el resultado de esta depuración.

Sobre explicar y comprender

En las páginas primeras del Libro I de la *Ética Nicomáquea* dejó escrito Aristóteles que "[…] no se ha de buscar el mismo rigor en todos los razonamientos" (1094*b*). Cerca ya de de las últimas procuraba aclarar las razones de su afirmación añadiendo que:

> Tampoco se ha de exigir la causa por igual en todas las cuestiones; pues en algunos casos es suficiente indicar bien el hecho, como cuando se trata de los principios, ya que el hecho es primero y principio. Y de los principios, unos se contemplan por inducción, otros por percepción, otros mediante cierto hábito, y otros de diversa manera (*Ética Nicomáquea* I: 1098*a*-1098*b*)

La tradición aristotélica acerca de los criterios que permiten decidir si una explicación dada resulta epistemológicamente respetable ha sido objeto de una cuidadosa y sumamente influyente reivindicación por parte de von Wright (1971), quien toma como punto de partida la arraigada contraposición entre explicaciones mecánicas, netamente causales, propias de la ciencia galileana, y explicaciones finalistas, irreductiblemente teleológicas, propias del aristotelismo. La impugnación de estas últimas, o el intento de reducirlas a las primeras, constituye como bien

señala von Wright (1971: 4) uno de los pilares del positivismo, junto con la doctrina de la unidad de la ciencia y el carácter canónico del modelo de explicación que ofrecen las ramas plenamente matematizadas de la física. Como no podía ser de otra forma, la restitución de la legitimidad de la explicación teleológica, y en general de la diversidad de modelos explicativos válidos, se convertiría así en uno de los ejes de los movimientos antipositivistas que ya desde finales del s. XIX acompañarían al auge del positivismo. Muy al inicio de aquellos forcejeos, Johann Gustav Droysen, a la sazón profesor de Historia en la Universidad de Jena, delineó en 1858 una "dicotomía metodológica" –la que diferencia entre explicación, *Erklären*, y comprensión, *Verstehen*– destinada a sustentar a su vez la distinción entre ciencias naturales y ciencias sociales, o del espíritu –asumiendo la explicación el papel de objetivo de las *Naturwissenchaften* y a la comprensión el de finalidad de las *Geisteswissenschaften*–; esta misma dicotomía, desplegada luego en manos de Wilhelm Dilthey, Max Weber o Robin G. Collingwood, acabaría por fraguar las bases sobre las que von Wright pretende consolidar ese ejercicio de desagravio epistemológico de las causas finales[243]. No ha de resultar inesperado, a la luz de los lazos que unen la crisis del positivismo lógico a la transición entre conductismo y cognitivismo –los cuales hemos procurado desentrañar *supra*–, que en buena parte del debate reciente sobre la autonomía de la explicación psicológica resuene, con mayor o menor nitidez, el eco de los planteamientos de von Wright. De hecho, es ya un lugar común en la historiografía aludir –como hace el propio von Wright (1971: 6) en relación a las ciencias sociales y conductuales en general– a la huella indeleble que las presiones cruzadas entre positivistas y antipositivistas parece haber dejado en la reflexión teórica de los psicólogos, y que cabe remontar a la enrarecida atmósfera que las querellas entre ilustrados y románticos dieron en su día a la génesis de la psicología como ciencia. Quizá ni siquiera sea preciso recordar que el propio Wilhelm Wundt dividiría la disciplina que él mismo había contribuido más que nadie a afianzar en dos ramas nítidamente diferenciadas: la *physiologischen Psychologie*, una ciencia natural basada en la introspección experimental y limitada a los fenómenos psíquicos de menor complejidad (a saber: en los que cabe mantener al margen la influencia del lenguaje y, con él, de la cultura), y la *Völkerpsychologie*, que permitiría abordar la inmensa complejidad de la mente humana siempre que se asumiera la necesidad de hacerlo mediante los métodos propios de las ciencias del espíritu.

[243] Antes aun del trabajo seminal de Droysen, como es sabido, hay referencias más bien imprecisas a la idea de comprensión en la obra de Giambattista Vico (1725), que, significativamente, surgen en un substrato anticartesiano: la reivindicación del conocimiento histórico frente a la primacía que la Ilustración, como hiciera Descartes, comenzaba ya otorgar a la mecánica –más en particular, a los los *Principia Mathematica Philosophiae Naturalis* (Newton, 1687)– en tanto que modelo de conocimiento pleno. Es habitual reconocer en Vico a un precursor del Romanticismo, como lo es ver en el Romanticismo –*cf.* por ejemplo, Leahey (2005: 165-170, 176-177)– una fuerza decisiva, *a contrario*, en la conformación de las aspiraciones de equiparar su estatus epistemológico al de las ciencias físicas que abrigaba la naciente psicología.

Pues bien: detenerse en el análisis de la idea misma de causalidad –un lugar al que la controversia sobre la autonomía explicativa de la psicología, como ha quedado bosquejado, se ha visto abocada no sin cierta reticencia– es –vamos a verlo enseguida– consustancial a los propósitos de von Wright, que bien pueden condensarse en el de mostrar que hay algunos usos importantes de la idea de causalidad, en relación con la de explicación, que no se ajustan al modelo de explicación nomológico-deductiva esquematizado por Hempel (1942, 1965) y que alcanzaría valor canónico en el seno del positivismo (von Wright 1971: 15). De hecho, su capacidad o incapacidad de dar cuenta de la explicación de la acción humana constituye, de acuerdo con el diagnóstico de von Wright (1971: 23), "la prueba final de la validez universal de la teoría de la explicación por subsunción" –es decir, del modelo nomológico-deductivo de Hempel, también conocido como modelo de cobertura legal[244].

Paso a paso: al presentar el aparato formal que sustenta su investigación, von Wright advierte sin titubear que los "estados de cosas" p_1, p_2, … p_n susceptibles de quedar enlazados por relaciones causales o, más en general, nómicas, han de ser estados lógicamente independientes entre sí y de carácter genérico, en el sentido de que "[…] puedan darse, o no, en ocasiones determinadas –y así darse, o no, repetidamente" (von Wright 1971: 43). Pero lo que en el contexto de esas indagaciones en torno a la explicación de la acción humana aparece guarnecido de un aire poco menos que axiomático –si se quiere, como un postulado del sentido común, o como una tautología cuya demostración resultaría ociosa–, es la constatación que, pese a encontrarse ya prefigurada en ciertos aspectos del trabajo de Skinner, terminaría por instaurar la distancia entre un fisicalismo de casos y uno de tipos y por abrir así camino a la corriente antirreduccionista que, de la mano del funcionalismo, más vigorosamente ha logrado penetrar en la reflexión contemporánea sobre la naturaleza de la explicación psicológica[245]. Es más: una defensa canónica de la autonomía de la psicología, como la ensayada por Pylyshyn (1984), bien puede como hemos visto acabar por recalar en la necesidad de concebir

[244] Es fácil hacer transparente el sentido de la expresión "cobertura legal": el caso particular que se pretende explicar –el color, digamos, de este o aquel cuervo– queda cubierto, o subsumido, bajo la ley general que enuncia que todos los cuervos son negros. El propio von Wright (1971: 11), no obstante, se ocupa de recordarnos que el término "covering-law model", que hizo fortuna en inglés, no se debe en realidad a Hempel sino a William H. Dray (1957), un historiador muy crítico con las tesis positivistas que insistió incansablemente en el carácter *sui generis* de la explicación de la acción humana (*cf*. von Wright 1971: 25).

[245] En efecto, ha quedado ya apuntado como la diferenciación entre las diversas instancias o casos de un estado mental y la clase o tipo de estado al que todas ellas pertenecen, de capital importancia en el seno del cognitivismo, parece enraizarse en las reflexiones de Skinner acerca de la naturaleza de estímulos y respuestas (Skinner 1935: 476-477; 1953: 104, *supra; cf.* también Gondra 1992: 41-42, 49, *infra*). El ardor con el que Skinner combatiría la "restauración cognitiva de la Casa Real de la Mente" (Skinner 1987: 784, *supra*) tiene evidentemente sus propios cauces, ajenos a la confluencia en este punto del análisis, al que, como se ve, se aviene también von Wright en su examen genérico de la idea de relación nómica.

también la idea de relación causal inscrita en nuestro modelo de explicación científica bajo la misma óptica de la distinción entre tipos e instancias[246]. Que dilucidar la cuestión de la autonomía de la psicología requiere en todo caso *alguna* relectura de nociones metafísicas fundamentales, como la de causalidad, es una conclusión que varias veces ha ido quedando hilvanada a lo largo de este trabajo.

De hecho, la distinción entre descripción y explicación en la que Pylyshyn (1984) fundamenta su defensa de la autonomía explicativa de la psicología comparte algunas vigas maestras con la distinción entre explicación y comprensión tal como la trazaba von Wright (1971). La clave de arco que sustenta la distinción en Pylyshyn es –recuérdese– la tesis de que el concepto de descripción es extensional y el de explicación intensional. Una explicación –piensa Pylyshyn– lo es en la medida en que captura determinadas generalizaciones, y "[…] diferentes vocabularios revelan diferentes generalizaciones" (Pylyshyn 1984: 2, *supra*). Es decir –recordemos–, que al contrario de lo que ocurre en la descripción de una secuencia de acontecimientos *X*, *Y*, el giro "La ocurrencia de *X* (dado un cierto conjunto de generalizaciones, Γ) explica la ocurrencia de *Y*" genera un contexto opaco, ya que que si *X* o *Y* son reemplazadas por términos coextensivos pero de sentido ajeno a Γ la sustitución no se hará *salva veritate*. Al afrontar el análisis del silogismo práctico como esquema básico de explicación psicológica, en la estela de Anscombe (1957), von Wright (1971: 121) observa que su validez formal requiere que "[…] the item of behavior mentioned in its conclusion is described (understood, interpreted) as action, as the doing or trying to do something by the agent under consideration". Así pues, añade, "[…i]n order to become *teleologically explicable* […] behavior must be first *intentionalistically understood*". La idea de que una cierta descripción de un fenómeno –o una cierta *comprensión* de un fenómeno: von Wright emplea ambos vocablos, como se ve, con idéntico sentido en este contexto– alumbre explicaciones que de otro modo quedarían veladas no puede menos que ofrecernos un aire de familiaridad. Dejando de lado el nítido eco del pensamiento de Max Weber –la acción, en tanto que "conducta subjetivamente significativa" posee un sentido interno (*Sinn*) en cuyo desentrañamiento consiste precisamente la compensión (*Verstehen*)[247]–, y por desenmadejar siquiera un hilo entre tantos: ya se ha comentado la insistencia de Pylyshyn (1984: 16-17, *supra*) en que "[…] theories address phenomena as 'events under descriptions'", lo que es tanto como decir que diferentes vocabularios descriptivos generan en realidad diferentes *explananda*. Así, por ejemplo, el vocabulario conductista nos fuerza –asegura Pylyshyn (1984: 8, *supra*)– a clasificar los eventos que intervienen en la explicación según una taxonomía basada en criterios físicos, mientras que el vocabulario cognitivista nos permite clasificarlos según criterios semánticos o intencionales[248]. O con otras palabras –salvando el hecho de

[246] *Cf*. Pylyshyn (1984: 11, *supra*), también Toribio (1991: 5, *infra*).

[247] *Cf*. por ejemplo, los ensayos de tema metodológico recogidos en Weber (1984).

[248] En su crítica del carácter subrepticiamente cognitivo del vocabulario explicativo del conductismo, de raíces chomskianas, Pylyshyn señala que a la hora de clasificar un movimiento como tal o cual respuesta "[…] only movements intended a certain way are counted" (Pylyshyn 1984: 8-9, *supra*).

que Pylyshyn evita ligar la explicación cognitiva a la noción tradicional de causas finales:

> […t]he *explanandum* of a teleological explanation is an action, that of causal explanation an intentionalistically noninterpreted item of behavior, *i.e.* some bodily movement or state. […*T*]*he same* item of behavior which is the *explanandum* of a causal explanation may also be given an intentionalistic interpretation which turns it into the *explanandum* of a teleological explanation. (von Wright 1971: 124)

Por diáfanos que resulten los vínculos entre sus concepciones de la explicación psicológica, no es difícil hallar más signos de la distancia entre Pylyshyn y von Wright. A veces, no obstante, tales signos son sólo aparentes. Así, por ejemplo, Pylyshyn parece cifrar la autonomía de la explicación psicológica en el propio ámbito del concepto de explicación –en la idea, en particular, de que distintos *vocabularios explicativos* permiten atrapar distintas generalizaciones–, y no en el de la descripción –diferentes descripciones, interpretaciones o comprensiones del mismo fenómeno, dice von Wright, alientan diferentes explicaciones. Pero lo que hace un vocabulario explicativo, según describe su trabajo el propio Pylyshyn, no es muy distinto de lo que hace una descripción a ojos de von Wright: si con cada acto de interpretación "[…] los hechos en cuestión quedan coligados bajo un […] concepto" (von Wright 1971: 135), un vocabulario refleja el modo en que dividimos "[…] un conjunto de alternativas", fija "[…] un trasfondo de distinciones y equivalencias" y entraña "[…] diferentes estructuras en el espacio de eventos posibles del que un evento observado en particular se toma como instancia" (Pylyshyn 1984: 16-17, *supra*) –el mismo tipo de labores que solemos encomendar precisamente a los conceptos[249]. Más abiertamente: cuando Pylyshyn evoca la implacable crítica chomskiana del trabajo de Skinner sobre conducta verbal, nos recuerda que el "vocabulario 'conductual'" de los estímulos y los reforzadores cosecha su eficacia explicativa del hecho de que, en la práctica e implícitamente, "[…] estas categorias son cognitivas" (Pylyshyn 1984: 8-9, *supra*), dejando claro que los términos de un vocabulario son –como en Locke– el trasunto verbal de determinadas categorías. En Pylyshyn como en von Wright, así pues, es la naturaleza de los conceptos empleados en la descripción de los fenómenos lo que determina que diferentes explicaciones, precisamente por sustentarse en diferentes descripciones, permitan rendir cuenta de diferentes hechos acerca de tales fenómenos

---

Parece seguirse, pues, que en el vocabulario explicativo del cognitivismo sucede exactamente eso: que asuntos tales como la intención son parte crucial de lo que determina la clasificación de un movimiento a efectos de su intervención en una teoría cognitiva. En efecto –asegura Pylyshyn *ibidem*–, la formulación de una teoría cognitiva requiere "[…] that we be capable of speaking of such things as […] the action intended by a certain behavior […]". Eso es *in nuce* lo que entiende von Wright por una explicación teleológica –aunque la idea de que el sujeto de la intención de una acción sea la conducta no resulte del todo precisa.

[249] De hecho, Pylyshyn utiliza esporádicamente la expresión "vocabulario descriptivo" como sinónimo –parece– de "vocabulario explicativo": *cf. e.g.*, Pylyshyn (1984: 16-17, *supra*). Ya se ha hecho notar *supra*, además, que "vocabulario" parece referirse en los argumentos de Pylyshyn más a un aparato conceptual que a un mero léxico.

o capturar diferentes generalizaciones[250]. Nada tiene de raro, entonces, constatar que el trabajo de Pylyshyn pueda leerse (*cf*. *supra*) como un cuestionamiento de la premisa, de aire hempeliano, según la cual el objeto de una explicación, o más en general de una disciplina científica, está naturalmente constituido con anterioridad al desarrollo de ésta –o, como solían decir los críticos del positivismo lógico, de que el vocabulario observacional de la ciencia es previo a su vocabulario teórico e independiente de él. Desde luego, no es mucha la distancia que nos separa ya del convencimiento de Davidson respecto a que las generalizaciones nómicas que puedan predicarse de determinados tipos de estados mentales –y, con ello, las relaciones causales en las que estos se involucren– dependen del hecho de que dichos tipos se forman bajo consideraciones normativas, que apelan a cánones de normatividad. Una vez más, la penetración de los argumentos antipositivistas en las raíces de la concepción cognitiva de la psicología es palpable.

Uno de los rasgos de la acción humana que de manera más notoria dificultan asimilar su explicación al modelo de Hempel es el papel que desempeñan convenciones y normas sociales en dicha explicación –o, más exactamente, en la tipificación de cada acción particular en géneros o clases susceptibles de convertirse en términos de generalizaciones legaliformes. Esta cuestión –que hemos escudriñado ya en relación con las observaciones de Pylyshyn (1984: 7, *supra*) acerca de lo sencillo que nos resulta predecir la conducta de una persona que tras contemplar un accidente marca los dos primeros dígitos del teléfono de emergencias, y que no sería difícil enlazar también con el pensamiento de Davidson– había tomado carta de naturaleza mucho antes en el seno de la crítica del positivismo, de la mano del estudio de Winch (1958) sobre la filosofía de la sociología. Tal como anota von Wright (1971: 28), el sociólogo "[…] debe comprender el 'significado' de los datos conductuales que registra para convertirlos en hechos sociales", lo que sólo puede lograrse "[…] describiendo (interpretando) los datos en términos de los conceptos y reglas que determinan la 'realidad social' de los agentes estudiados"[251]. La necesidad de que el observador emplee en la explicación el mismo marco conceptual que rige en el fenómeno explicado conduce a Winch y a von Wright, que tienen bien presente la diversidad de tales marcos conceptuales, a una reflexión sobre la idea de observación participante y, con ello, a una muy matizada reivindicación de la doctrina que relaciona la comprensión, por oposición a la explicación, con la empatía. Aunque toda esa reflexión, y las capitales consecuencias de orden moral y político

[250] El hecho de que puedan coexistir dos descripciones verídicas de una misma secuencia de fenómenos de las que sólo una tenga valor explicativo, que Pylyshyn consigna al distinguir la extensionalidad de la descripción de la intensionalidad de la explicación, no es óbice –adviértase– para que la capacidad que distintas explicaciones atesoran de iluminar distintas generalizaciones provenga de su articulación sobre distintos esquemas conceptuales, llámense vocabularios explicativos, descripciones, interpretaciones o comprensiones.

[251] *Cf*. Pike (1967, *supra*): así como es nítido, según reconoce el propio Pylyshyn (1984: 13, *supra*), el eco de la distinción trazada por Pike entre propiedades *emic* y propiedades *etic* de la conducta, también lo son las resonancias recíprocas entre las propuestas de Winch y Pike, aunque ninguno de los dos cita al otro.

que acarrea, apenas pueden entreverse implícitamente en Pylyshyn[252], ni por lo general en otros adalides de la ortodoxia cognitivista, es difícil ver como podrían ser esquivadas una vez que se ha admitido la relevancia de lo convencional –de *nómos*, por oposición a *phýsis*– de cara a la tarea de clasificación de las conductas humanas que se halla indisolublemente ligada a su explicación. El matiz subjetivista que al menos desde Dilthey (1883) ha venido tiñiendo la idea de comprensión, hermanándola a unas dotes empáticas de naturaleza más o menos trascendente, ha contribuido sin duda al cerrado rechazo que dicha idea ha suscitado en quienes aspiraban a recabar para la psicología el respeto de una comunidad científica preponderantemente afín a planteamientos positivistas –siquiera vagamente positivistas[253]. Ahora bien, que convenir en que la clasificación de la conducta según

---

[252] Como cuando a vuelapluma deja escrito Pylyshyn, por ejemplo, que:

> […] what we see a stimulus *as* depends on what we know. And, of course, it is what we see things as that determines their effect on our behavior. When we see a certain red light *as a traffic light*, we stop –and our seeing the light as a traffic light depends on knowledge of our culture's conventions. That, in a nutshell, is why the physical properties of stimuli do not determine our behavior. (Pylyshyn 1984: 15)

Como está mandado en un cognitivismo de corte clásico, parece pasarse por alto la mera posibilidad de que existan circunstancias en que sean ciertas convenciones culturales, y no nuestro conocimiento de ellas, las que condicionen el efecto de determinados estímulos sobre nuestra conducta. Vagamente se apunta a algo parecido poco después, cuando se afirma que la delimitación de las regularidades conductuales que consideramos dignas de investigación está ella misma impregnada por nuestra cultura (Pylyshyn 1984: 17) –en formas, parece sobreentenderse, que el sujeto de la conducta no forzosamente conoce. En cualquier caso, la observación que con ese restringido alcanzado plantea Pylyshyn basta para sus propósitos.

[253] Como atinadamente subraya von Wright (1971: 6, 26), quien rechaza también tajantemente que la idea de comprensión haya de quedar por fuerza ligada a la de empatía (von Wright 1971: 30), ese aire subjetivista es vívido, por ejemplo, en Simmel (1892, 1918) o en Dray (1957). Ya Collingwood (1946) había intentado desproveer a la comprensión del matiz subjevista de Dilthey, al argumentar que los procesos psicológicos que se ponen en juego en la comprensión no son en realidad los del sujeto de la acción histórica, sino los que el historiador le atribuye a la luz de la evidencia disponible: se trata pues de reconstruir más que de revivir.

Es obligado señalar, sin embargo, que en Dilthey parecen convivir esa interpretación subjetivista –de "psicologista" la tilda González (2000: 127)– con otra de corte más hermenéutico en la que lo que ocurre en la comprensión no es que revivamos bajo el prisma de nuestra propia experiencia los motivos de la acción que comprendemos, sino que, al margen de esos motivos u otros procesos psicológicos que subyazgan, se revive "[…] el significado de una expresión vital", reconstruyendo "[…] una totalidad en cuyo seno se determina el significado de cada una de las partes" para comprender así "[…] una configuración espiritual que posee su propia estructura" (González 2000: 128). Bien podría discutirse –qué duda cabe– si tal lectura hermenéutica del concepto de comprensión contradice, y si lo hace con robustez, la idea de que la exuberante floración de maneras de entender dicho concepto que alberga la tradición antipositivista viene coaligada –como sostiene López de la Vieja (2009)– por el hecho de que en todos los casos se "[…] asocia la inteligibilidad del objeto investigado a las experiencias subjetivas"; es decir, si puede verdaderamente erguirse una noción hermenéutica de comprensión firme y ajena a todo psicologismo. Así, la idea primigenia de comprensión tal como queda articulada en Droysen resulta a ojos de González (2000) más afín a la interpretación hermenéutica del pensamiento de Dilthey que a su interpretación psicologista; López

criterios psicológicos es impractible si queda desprovista de la referencia a determinados elementos de índole convencional nos aboque a una ineludible reflexión metodológica no entraña que nos aboque a asumir ninguna concepción en particular de nuestra capacidad empática[254], ni tampoco –valga decir– una comprensión subjetivista de la comprensión.

Las objeciones de von Wright al modelo de explicación de Hempel, con todo, parten de una constatación intuitiva que no depende de que la acción humana sea efectivamente *rara avis* –ni mucho menos de que, como veremos que von Wright reclama, su comprensión sea previa a nuestra idea de causalidad. A saber: afirmar que un pájaro es negro porque es un cuervo y todos los cuervos lo son es una explicación que se ciñe impecablemente al esquema de subsunción desarrollado por Hempel, pero respecto a la que "instintivamente" dudaríamos de que constituya una auténtica *explicación*[255]. Antes bien:

> What is required, if our search for an explanation is to be satisfied, is that the basis of the explanation be somehow more strongly related to the object of explanation than simply by the law stating the universal concomitance of the two characteristics [ravenhood and blackness]. (von Wright 1971: 19)

Esa mayor fortaleza puede buscarse –apunta von Wright– por dos caminos: la indagación empírica sobre las causas eficientes de la relación entre las dos propiedades cuya concomitancia describe la ley en cuestión –sobre su fundamento categórico, si se prefiere: la característica compartida por todos los cuervos que es directamente responsable del color de su plumaje–, o bien, a modo de *fiat*, la concesión a dicha relación del rango de elemento definitorio de una de las propiedades involucradas –en el ejemplo, considerar el color negro como condición de pertenencia al género *Corvus*, o a algunas de las especies que comprende si es que a ellas se refería específicamente la ley, en el sentido de admitirlo como integrante del significado del término que designa a la clase natural en cuestión. *Mutatis mutandis*:

---

de la Vieja (2009), en cambio, compendia la idea de comprensión de Droysen como "[…] una suma peculiar de elementos cognitivos, evaluativos y expresivos" en la que "[…l]a subjetividad era incorporada como un elemento relevante para la investigación". Salta a la vista, en todo caso, que si la mera mención de la empatía como condición del conocimiento ha podido suscitar recelos en los círculos más afines a la tradición positivista, expresiones vitales y configuraciones espirituales no habían de cosechar una adhesión más entusiasta.

[254] La viabilidad de una explicación de la empatía en términos de procesos de razonamiento es de hecho cuestión muy debatida en el seno de la psicología cognitiva, donde toma la forma de un vivo debate entre quienes abogan por la tesis de que la empatía es *sensu stricto* una *teoría* de la mente –un conjunto de hipótesis que construimos acerca de las creencias y deseos de los otros– y quienes se inclinan por considerarla un fenómeno de *simulación* –una habilidad independiente del razonamiento teórico que nos permite, como suele decirse, ponernos en la piel del otro. Ni una parte ni otra, dicho sea de paso, parece inclinada a considerar la empatía un fenómeno refractario a la explicación. Puede encontrarse una apasionante revisión de esta controversia en Carruthers y Smith (1996).

[255] *Cf.* el examen de la reivindicación de las explicaciones disposicionales ensayado por Place (1999: 383-384) contra las acusaciones de circularidad de Geach (1957), que quedó desplegado *supra*.

hacer descansar las perspectivas del cognitivismo sobre la identificación empírica de los estados psicológicos que explican ciertas conductas –entendidos no como meras disposiciones sino como su sustrato categórico, por mucho que su clasificación en tipos psicológicos se haga depender de criterios funcionales y no físicos–, o bien hacerlas descansar sobre el análisis del significado del vocabulario psicológico ordinario, han venido siendo las dos rutas cardinales de desarrollo de la concepción funcionalista de lo mental. Una certera anotación de von Wright (1971: 19) apunta a un ámbito de acuerdo en el seno de las distintas vertientes del funcionalismo: sea cual sea la estrategia elegida, la concomitancia entre las propiedades mencionadas en la ley habrá de tornarse no sólo universal, sino también, en un sentido u otro, necesaria[256] -incluso, como hemos visto al hilvanar la controversia sobre el dolor y el jade (*supra*), en aquellas variedades del funcionalismo en las que, con Fodor, se entiende la relación entre estados mentales y estados funcionales de un sistema como una cuestión sujeta a descubrimiento empírico, no a dilucidación conceptual.

En cualquier caso, ni siquiera un terminante reconocimiento de que la explicación de la acción humana, por inherentemente teleológica, sea irreducible a un modelo de explicación ceñido a la especificación de causas estrictamente mecánicas zanjaría la cuestión a ojos de von Wright. Antes bien, ése sería para él nada más que el lugar desde el que emprender el paso decisivo, mucho más osado, que vendría dado por la constatación de que:

> […] we cannot understand causation, nor the distinction between nomic connections and accidental uniformities of nature, without resorting to ideas about doing things and intentionally interfering with the cause of nature. (von Wright 1971: 65-66)

Sería entonces la noción de causa eficiente la que reposaría sobre la de causa final, el mecanismo sobre la teleología y no al contrario –si bien von Wright (1971: 74) acaba por conceder que no cabe modo de dirimir si la noción de acción es previa a la de causa o viceversa, por mucho que la de acción parezca siempre, a su juicio, llevar la delantera. Así, *p* resultaría ser la causa de *q* si y sólo si provocando (*intencionalmente*) la ocurrencia de *p podríamos* provocar la de *q*, o evitando aquella eludir ésta (von Wright 1971: 70): ésa y no otra sería nuestra noción de causalidad –una que, como es patente, se halla profundamente impregnada no sólo de la de acción intencional sino también de la de posibilidad[257].

---

[256] Que "[…] el sello de la conexión nómica, de la legaliformidad, es *la necesidad y no la universalidad*" es además, a juicio de von Wright 1971: 22), una conclusión ineludible del análisis de los enunciados contrafácticos inspirado por los trabajos de Chisholm (1947) y Goodman (1947), que nos obliga a dotarnos de herramientas para distinguir una correlación perfecta entre *p* y *q* –una "concomitancia universal 'accidental'", en palabras de von Wright– de una conexión causal de naturaleza nómica entre *p* y *q*, en la medida en que la segunda, pero no la primera, se nos antoja como un fundamento sólido para nuestra convicción de que si no hubiera ocurrido *p* tampoco habría ocurrido *q*. La idea de necesidad nos proporcionaría precisamente tales herramientas.

[257] No menos enérgico, desde luego, es el giro que von Wright pretender imprimir a nuestra concepción de las relaciones entre la acción voluntaria y su substrato neurológico, en líneas de intenso

En definitiva, un replanteamiento de la idea de causalidad, éste sin duda de naturaleza mucho más radical, se atisba de nuevo en el estuario en el que confluyen las diversas reivindicaciones de la autonomía de la explicación psicológica que venimos escrutando: hemos de reconstruir nuestra noción de causalidad –nos dice von Wright– de modo que se hagan transparentes ciertas vetas que la recorren, en las que se hallan cristalizadas las ideas de posibilidad, necesidad y acción intencional. No está de más, tal vez, rememorar en relación con esto la conclusión que Smith (2002b: 242) alcanzara al término de su dialogo con Dennett (2002) en torno a las perspectivas de naturalización del concepto de intencionalidad: "Because ontological categories are in part intentionally constituted, attempting to explain representation while dining out on ontology is […] fatally circular". Bien parecería que la misma intuición que alienta el diagnóstico de Smith subyace al hecho de que Pylyshyn (1984) eligiera para cerrar su estudio de los fundamentos de la ciencia cognitiva un párrafo de William James (1892: 467) en el que se asegura que la idea de psicología como ciencia natural no apunta, como querría el positivista, hacia:

> […] a sort of psychology that stands at last on solid ground. It means just the reverse; it means a psychology particularly fragile, and into which the waters of metaphysical criticism leak at every joint, a psychology all of whose elementary assumptions and data must be reconsidered in wider connections and translated into other terms.

Una psicología –vale la pena repetirlo– por cuyos goznes se filtran constantemente las aguas de la crítica metafísica.

---

aire cartesiano que parecen una consignación *avant la lettre* de las mismas preocupaciones a las Libet (1985) –*cf*. también Libet, Gleason, Wright y Pearl (1983)– daría una respuesta bien distinta, aunque igualmente polémica. Dice von Wright (1971: 77):

> An example of a basic action could be the raising of (one of) my arm(s). Suppose one could "watch," one way or another, what happens in my brain and that one has been able to identify the neural event, or set of events, *N*, which must occur, we think, if my arm is to rise […]. I say to somebody: "I can bring about the event *N* in my brain. Look." Then I raise my arm and my interlocutor observes what happens in my brain. He sees *N* happen. But if he also observes what I do, he will find that this takes place a fraction of a second after N. Strictly speaking: what he will observe is that the result of my action, *i.e.*, my arm going up, materializes a little later than *N* occurs.
>
> This is causation operating from the present towards the past. It must, I think, be accepted as such. By performing basic action we bring about earlier events in our neural system.

El propio von Wright (1977: 191), por lo demás, anota la cercanía de su propuesta a la adelantada por Chisholm (1966), si bien rechaza por innecesariamente oscuro el concepto de *causación inmanente* que Chisholm construye para intentar dar cuenta de la relación entre las acciones de un sujeto y los acontecimientos mundanos.

ENTRE EL MUNDO Y LA MENTE: LITIGIOS FRONTERIZOS

Los lazos con el mundo: cómo describir estímulos y respuestas

Es difícil que pase inadvertido el hecho de en que los modelos funcionalistas de la naturaleza de lo mental se acostumbra a restringir el formalismo de descripción funcional a los estados internos: se trata canónicamente, como hemos visto, de la explicitación del correlato funcional del estado mental en cuestión mediante un enunciado de Ramsey, a veces denominada explicitación del *correlato Ramsey-funcional* del estado mental (Block 1978: 67). No se acomete, en cambio, la caracterización funcional de las aferencias y eferencias que, junto con esos estados internos, conforman la descripción global del sistema estudiado –ya sea su tabla de máquina u otra forma de especificar su arquitectura y procesos computacionales; ya con base en la psicología ordinaria o un fragmento articulado y significativo de ella, ya en el corpus teórico de una psicología científica acabada.

Así, por ejemplo, en el autómata simplicísimo –capaz tan sólo de dictaminar si el número de estímulos iguales que ha recibido es par o impar– que Block (1996) presenta para ejemplificar las tesis funcionalistas, sus dos estados internos posibles se definen por sus relaciones recíprocas, con el único estímulo admitido, "1", y con las dos respuestas posibles, "Par" o "Impar"; la caracterización del estímulo y de las respuestas, en cambio, ignora las relaciones que estos establezcan con los estados internos. En un caso como éste, podríamos desde luego tomarnos el llevadero trabajo de definir –pongamos por caso– la emisión de la respuesta "Impar" como:

Emitir "Impar" = Ser un x tal que $\exists R \exists S$ [Si x emite R x ha recibido la aferencia "1" y se encuentra en el estado "S_1"; la emisión de R va acompañada del paso al estado "S_2"; si x emite S x ha recibido la aferencia "1" y se encuentra en el estado "S_2"; la emisión de S va acompañada del paso al estado "S_1" & x emite R].

No es aventurado concluir sin más que una maniobra análoga, más sencilla si cabe, podría aplicarse a la definición relacional del estímulo "1" –así como, claro está, de la respuesta "Par". De la misma manera, la tabla de máquina que describe el funcionamiento de un viejo expendedor automático de refrescos (Nelson 1975: 252; Block 1978: 66, 1980: 30) –convertida con el paso de los años en el ejemplo paradigmático, en el sentido acuñado por Kuhn 1962, de redescripción como máquina de Turing de un sistema que ejecuta una tarea según un procedimiento efectivo– fija las aferencias admitidas por el aparato (monedas de cinco o diez centavos, en el ejemplo original) y sus posibles eferencias (un refresco, una moneda de cinco centavos) *tal cual*. No se plantea la caracterización de "moneda de diez

centavos" como aferencia tal que si el sistema se encuentra en el estado S_1 provoca la expedición de un refresco y la permanencia del sistema en ese estado, y si el sistema se encuentra en el estado S_2 provoca la expedición de un refresco y de una moneda de cinco centavos, así como la transición a S_1.

Evidentemente, "moneda de diez centavos" *es* una descripción abstracta con respecto a las propiedades físicas de una pequeña pieza de metal, aunque sea sólo en la medida en que prescinde de determinadas diferencias en dichas propiedades que son irrelevantes en relación con el valor de la moneda. Cabe argumentar, incluso, que es una descripción *funcional*, que agrupa a ciertas piezas de metal bajo la clase "moneda de diez centavos" en virtud de *algunos* de sus efectos (entre ellos, precisamente, *algunos* de sus efectos sobre el funcionamiento de máquinas expendedoras de refrescos), y que es además *convencional* en un sentido obvio: si la autoridad monetaria competente decretara que se admitieran como monedas de diez centavos piezas físicamente diferentes, éstas pasarían a ser monedas de diez centavos. Pero parece claro que la simplificación efectuada al describir las aferencias del expendedor de refrescos en esos términos tiene únicamente intención didáctica, y que en un desarrollo pleno de la tabla de máquina en cuestión, las descripciones "moneda de diez centavos" y "moneda de cinco centavos" serían reemplazadas por descripciones estrictamente físicas, las cuales pasarían seguramente por una conjunción de valores dentro de cierto rango en varias magnitudes físicas –diámetro, grosor, peso, etc.

Reiteradamente deja claro Block (1978) que no está en el ánimo de la concepción funcionalista de la mente –no, al menos, por regla general– la caracterización funcional de estímulos y respuestas, sino su caracterización física[258]:

> Functionalists have tended to treat the mental-state terms in a functional characterization of a mental state quite differently from the input and output terms. […]
> […F]unctionalism […] has typically insisted that characterization of mental states should contain descriptions of inputs and outputs in *physical* language. (Block 1978: 64)

Aunque no es cierto –como hemos de ver– que dicha insistencia en la descripción física de estímulos y respuestas sea común a todas las escuelas y corrientes que conviven bajo la inspiración funcionalista, su expresión más rotunda –que el propio Block registra– acaso sea la que en consonancia con sus estrictos votos materialistas le otorgara Armstrong (1968):

> We may distinguish between "physical behaviour," which refers to any merely physical action or passion of the body, and "behaviour proper," which implies relationships to the mind […] Now, if in our formula "behaviour" were to mean "behaviour proper," then we would be giving an account of mental concepts in terms of a concept that already presupposes mentality, which would be circular. So it is clear than in our formula, "behaviour" must mean "physical behaviour." (Armstrong 1968: 84)

[258] Dicho de otro modo: el funcionalismo se alinea en este terreno de lado de Watson, contra Skinner: *cf*. Mandler (2002: 341, *supra*).

La transparencia del razonamiento de Armstrong hace patente, además, cuál es la motivación cardinal que subyace al rechazo de la descripción funcional de estímulos y respuestas: tal estrategia arrojaría una explicación circular, del estilo de la *vis prolifica* o la *virtus dormitiva*. Gracias a una especificación de estímulos y respuestas libre de los matices mentalistas que le daríamos al establecerla en términos funcionales, en cambio, la concepción funcionalista de la mente podría –como plantea Block (1978)– quedar de algún modo trabada con la realidad, y evitaríamos que acabara girando en el vacío:

> Therefore, functionalism can be said to "tack down" mental states only at the periphery –i.e., through physical, or at least nonmental, specification of inputs and outputs. (Block 1978: 64)

Pero las tribulaciones que llevan a Block a rechazar la caracterización funcional de aferencias y eferencias tal vez tengan que ver, más aún que con el temor a incurrir en un esquema explicativo irremediablemente circular, con el de que tal caracterización nos empuje a conceder carta de naturaleza mental con demasiada prodigalidad[259]. Si en el marco de una teoría funcionalista de la mente concediéramos que el vocabulario observacional en que describiésemos estímulos y respuestas no quedara anclado a descripciones físicas categóricas, sino que lo dejáramos teñirse de los mismos procedimientos de definición funcional que hubiéramos articulado para los estados internos, entonces –apunta Block–,

> […] a system could be functionally equivalent to you if it had a set of states, inputs and outputs causally related to one another in the way yours are, no matter what the states, inputs and outputs were like. (Block 1978: 90)

He aquí, entonces, el fantasma del liberalismo que parece atormentar a Block: la desmesura en la atribución de estados mentales que vendría propiciada por el levantamiento de aranceles que, por así decir, supondría la caracterización funcional de estímulos y respuestas. En efecto, ya en los párrafos iniciales de ese trabajo –uno de los más tempranamente dedicados a advertir de los resquicios que, en cualquiera de sus vertientes, debilitaban la concepción funcionalista de la mente– Block proclamaba sin clemencia su tajante y desalentadora conclusión: "[…] the troubles ascribed by functionalism to behaviorism and physicalism infect functionalism as well" (Block 1978: 63)[260]. Cuando el funcionalismo se desmorona es –piensa Block– cuando ha de enfrentarse a la tarea de decidir a qué sistemas está justificado atribuir estados mentales y a cuáles no: cualquier formulación de las tesis funcionalistas que podamos construir será o bien demasiado liberal, otorgando vida mental a entidades

[259] Cf. Block (1996: 24, *supra*), al hilo de las raíces que hunde el funcionalismo en el pensamiento de Hilbert y su polémica con Frege.

[260] De dicha conclusión, dicho sea de paso, el propio Block se retractaría parcialmente más tarde: *cf.* Block (1986: 100).

a las que nuestras intuiciones nos dicen que es absurdo concedérsela, o bien demasiado chauvinista, negándosela a entidades a las que esas mismas intuiciones nos dicen que deberíamos otorgársela[261]. En suma, la estacada que según Block ataja el paso al funcionalismo puede quedar perfilada –así, de hecho, lo plantea Bechtel (1988: 177)– como el fracaso en su empeño fundacional de mantener una escrupulosa equidistancia entre conductismo y fisicalismo.

> Al igual que el conductismo filosófico, [el funcionalismo] apela a criterios conductuales para caracterizar los estados mentales, pero, a diferencia de él, […] interpreta los estados mentales como estados internos y les otorga un papel causal en la producción de la conducta. Al apoyar [la consideración de] los estados mentales como procesos internos, el Funcionalismo está de acuerdo con la Teoría de la Identidad, pero difiere [de ésta] en que no insiste en [que los] tipos de estados mentales se identifican con [tipos de] estados del cerebro. (Bechtel 1988: 177)

El funcionalismo se vería entonces obligado a dejarse vencer ya del lado del fisicalismo, ya del lado del conductismo, y a asumir la carga de problemas que corresponda:

> O el Funcionalismo será como el conductismo filosófico al ser demasiado liberal atribuyendo estados mentales a sistemas a los que no deberían atribuirse, o será, como la Teoría de Identidad, demasiado chauvinista al negar estados mentales a sistemas que los tienen. (Bechtel 1988: 177)

Pero lo que es seguro, si los argumentos de Block son sólidos, es que el funcionalismo sucumbirá a una u otra iniquidad, que la aspiración de que el empleo de criterios funcionales llegue a habilitarnos para atribuir estados mentales –por así decir– *salomónicamente* no es sino una quimera.

Pues bien, el dilema cuyos cuernos son el chauvinismo y el liberalismo se extendería igualmente, según argumenta Block (1996), a la elección de un vocabulario para la descripción de los estímulos y las respuestas:

> If we characterize inputs and outputs in a way appropiate to our bodies, we chauvinistically exclude creatures whose interface with the world is very different from ours […]. The obvious alternative of characterizing inputs and outputs themselves functionally would appear to […] fall to the opposite problem of liberalism. (Block 1996: 23-24)

Es innegable, desde luego, que el talante liberal que parece connatural al funcionalismo se vería acentuado en la medida en que, bajo descripciones

[261] La pregunta por el vocabulario idóneo para la caracterización de estímulos y respuestas, así pues, se torna particularmente determinante en el contexto de las primigenias preocupaciones de Turing (1950a, *supra*) respecto a la atribución de estados mentales a sistemas no humanos –de inteligencia, en su caso, a máquinas computadoras.

funcionales, dos estímulos o dos respuestas físicamente muy diferentes puedan pertenecer al mismo tipo de estímulo o de respuesta tanto si se trata de máquinas como de espíritus incorpóreos, de criaturas de distintas especies o de la misma, o de la misma criatura en distintos momentos. Así, por ejemplo, Block encara la caracterización funcional de estímulos y respuestas con la inquietud de que:

> […] characterizing inputs and outputs themselves functionally would appear to yield an abstract structure that might be satisfied by, for instance, the economy of Bolivia under manipulation by a wealthy eccentric [...] (Block 1996: 24)

Pero eso estaba ya –como quien dice– descontado en los escenarios que preocupan a Block en 1978, en los que un sistema al que somos vivamente reticentes a otorgar estados mentales resulta ser funcionalmente indistinguible de nosotros. La razón es que tales casos están cuidadosamente construidos para que las aferencias al sistema sean del mismo tipo que los estímulos a los que nosotros somos sensibles –se presta acaso algo menos de atención a las eferencias. En sus experimentos imaginarios sobre robots homunculares, por ejemplo, Block (1978: 70-74) no sustituye la totalidad del sistema nervioso de un ser humano por un ejército de homúnculos: deja intactos los órganos sensoriales con sus transductores, y el sistema musculo-esquelético con sus efectores. Pero la trampa es patente: si, *ex hypothesi*, el robot homuncular implementa una tabla de máquina que constituye una descripción completa del funcionamiento de un sistema cognitivo humano, no hay razón para excluir sus funciones perceptivas y motoras del experimento. El único motivo para hacerlo es desembarazarse de antemano de la posibilidad de que las señales que ingresan al sistema fueran también diminutas bombillas encendidas o apagadas, o cualquier otra cosa por el estilo, a la que los homúnculos encargados de la transducción sensorial pudieran responder a su manera, tan individualmente obtusa como colectivamente sagaz –*ídem* para las respuestas y los homúnculos efectores. Por supuesto, extender la jurisdicción de los homúnculos hasta la percepción y la conducta nos obligaría a adoptar una caracterización funcional de estímulos y respuestas *o* a renunciar a la premisa de que el robot homuncular implementa la misma tabla de máquina que la cobaya humana, y con ello, a la motivación teórica del experimento imaginario. Sin una caracterización funcional de estímulos y respuestas, no habría justificación para sostener –por ejemplo– que la concentración de determinadas moléculas en los receptores olfativos del sujeto es el mismo estímulo –un estímulo del mismo tipo, mejor dicho– que determinado patrón luminoso del panel de bombillas de su robot homuncular. Con la arbitraria decisión de desistir de la laboriosa sustitución de neuronas por homúnculos cuando llega a los nervios perceptivos y motores, Block no hace –en definitiva– sino tratar de desembarazarse de la necesidad de tomar otra decisión igualmente arbitraria, pero más patentemente ilícita: la de estipular qué cuenta como estímulos o respuestas del mismo tipo. De hecho, a la hora de afrontar la objeción de que dada la diferente naturaleza física del sistema homuncular éste quedaría sujeto a la interferencia de estímulos de naturaleza también diferente, Block

se muestra más tajante en cuanto al modo de descripción y clasificación de estímulos que ha adoptado: "[…t]he person who says what system he or she is talking about gets to say what signals count as inputs and outputs" (Block 1978: 72). Es decir: la caracterización de estímulos y respuestas que se sigue no es de índole funcional, sino estipulativa –o, si se prefiere, arbitraria[262].

En realidad, apreciar que la concepción funcionalista de los estados mentales conduce naturalmente a la caracterización funcional de estímulos y respuestas no requiere el concurso de concienzudos ejercicios imaginativos. Incluso en el caso elemental que Block (1996) elige para ilustrar la noción de tabla de máquina –el autómata que determina la cardinalidad par o impar del total de estímulos iguales que recibe– cabe desenvolver el mismo argumento que para el robot homuncular Si repudiamos la caracterización funcional del único estímulo posible del autómata, "1", y de sus dos respuestas, "Par" e "Impar", recién esbozada, entonces tendremos que rechazar también que un autómata cuyos estados de tabla de máquina vengan definidos por idénticas relaciones con el estímulo "·" y con las respuestas "El número de estímulos procesado dividido entre dos es un número entero" y "El número de estímulos procesado dividido entre dos no es un número entero" sea equivalente al detector serial de paridad, pese a que resulta evidente que lo es. En cambio, si admitimos caracterizar en términos funcionales los estímulos y respuestas del autómata, entonces infinitud de sistemas actuales o posibles serán detectores seriales de paridad –entre ellos, también, todo tipo de robots homunculares, así como la economía boliviana sometida a las caprichosas operaciones financieras de un extravagante potentado. Así que el conflicto entre liberalismo y chauvinismo se da ya entre los autómatas más humildes, y también en el modesto suburbio que habitan inclinarnos por una descripción física o funcional de sus lacónicas respuestas y sus exiguos estímulos tiene consecuencias que no es de recibo eludir estipulando a nuestro antojo los criterios taxonómicos que convengan a nuestros intereses teóricos puntuales. No podía ser de otro modo –al fin y al cabo–, dado que la fabulosa tabla de máquina que, *ex hypothesi*, proporciona la descripción funcional completa del sistema nervioso que ha de replicar el robot homuncular no se diferencia sino en su extensión de la que guía los pasos del tedioso detector serial de paridad. En definitiva, la desmedida liberalidad que Block imputa al proyecto de descripción funcional de estímulos y respuestas era ya un rasgo de la concepción funcionalista de la mente *per se*, con independencia del modo de inventariar los episodios de interacción con el entorno que pudiera adoptar.

Otro indicio de que la preocupación de Block al respecto es lluvia sobre mojado proviene de la propia forma de su descorazonado dictamen acerca de la viabilidad de una descripción de estímulos y respuestas que esquive tanto el liberalismo como

[262] Sin embargo, al rechazar uno de algunos ejemplos de tales interferencias físicamente dispares, como pudiera ser una riada del Yangtsé, Block (1978: 71) los compara con el caso en que una bomba impide que una computadora ejecute "[…] la tabla de máquina que fue construida para implementar": se engarzan así subrepticiamente en su concepción de lo mental las consideraciones teleológicas que otros, como Lycan (1987, *supra*), quieren hacer explícitas.

el chauvinismo, un calco del que él mismo ofrecía en cuanto a la posibilidad de que idéntica doble condición quede satisfecha por cualquier concepción funcionalista de los estados mentales. En efecto a la rotundidad con la que Block (1978: 65) aseguraba que "[…] indeed, any version of functionalism that avoids liberalism falls, like physicalism, into chauvinism" corresponde una franqueza pareja al trazar la conclusión relativa al vocabulario en que hayamos de describir estímulos y respuestas:

> The question is: is there a description of inputs and outputs specific enough to avoid liberalism, yet general enough to avoid chauvinism? I doubt that there is.
>
> Every proposal for a description of inputs and outputs I have seen or thought of is guilty of either liberalism or chauvinism. (Block 1978: 91).

No menos significativo, en este sentido, resulta el hecho de que los argumentos con los que Block condena por chauvinista la descripción categórica de estímulos y respuestas en la que coinciden funcionalistas empíricos y analíticos sean los mismos de los que se valiera para refutar el fisicalismo de tipos –o la interpretación del funcionalismo como una variante del fisicalismo de tipos preferida generalemente, como hemos visto, por los funcionalistas de corte analítico. Como el fisicalista de tipos, el funcionalista acorralado por la acusación de chauvinismo puede, en la estela de Lewis (1969), apelar a vastas construcciones disyuntivas: Lycan (1981), por ejemplo, asegura que se conformaría con que el funcionalismo pusiera en claro un conjunto de condiciones suficientes para la existencia de propiedades psicológicas, dejando la especificación de condiciones necesarias en manos de un enunciado disyuntivo que enumere los conjuntos de condiciones suficientes adecuados para que se den tales o cuales estados psicológicos en tales o cuales tipos de sistemas físicos[263]. Pero esa táctica –piensa Block– incurre en la misma debilidad que aqueja al propio fisicalismo de tipos: ante la pregunta por la naturaleza común de los fenómenos englobados por la disyunción –*qué es* un estado mental, qué un estímulo o una respuesta de tal o cual tipo–, sólo le quedará el balbuceo. El problema de fondo, tal como lo ve Block, es que especificar condiciones suficientes para la existencia de propiedades psicológicas deja en el aire la cuestión de cuál es la naturaleza última de esas propiedades –su naturaleza metafísica, diría Block (2007b: 9)–, en el sentido de que no aclararía la razón por la que podrían ostentar propiedades psicológicas, incluso propiedades psicológicas del mismo tipo, entidades que no satisficieran las mismas condiciones suficientes. La maniobra de asimilación del chauvinismo emprendida por Lycan, entonces, no nos llevaría todo lo lejos que nuestra legítima ambición explicativa quiere llegar. En particular, no nos ayudaría a entender por qué

[263] *Cf.*, *supra*, las críticas de Block y Fodor (1972a) a esta estrategia en la que, como se mencionó, Lycan confluye en buena medida con las propuestas de, entre otros muchos, Kim (1972, 1989, 1992).

tanto nosotros como sistemas físicamente (fisiológicamente, anatómicamente, etc.) muy diferentes podemos albergar estados mentales[264].

Además, si la amenaza del chauvinismo nos forzara a abrir el paso a una respuesta al problema de la caracterización de estímulos y respuestas en términos de disyunciones de vocabularios categóricos, la vecindad lógica entre esa cuestión y la de la propia naturaleza de los estados mentales acabaría irremisiblemente arrastrando al funcionalismo hacia el fisicalismo. Como bien dice el propio Block (1978: 91), "[…i]f functionalists suddenly smile on wildly disjunctive states to save themselves from chauvinism, they will have no way of defending themselves from physicalism".

Lo que todo esto pone de manifiesto es, de nuevo, que articular un aparato funcional de descripción de aferencias y eferencias no empeora, *pace* Block, la posición del funcionalismo ante el dilema que él mismo pergeñara. Inclinarse por un vocabulario funcional para la especificación de estímulos y respuestas –es cierto– obliga al teórico funcionalista a adentrarse en la *terra incognita* del liberalismo en cuanto concierne a la delicada tarea de otorgar estados mentales a otros. Pero eso es perfectamente irrelevante, porque lo único que lo retenía cerca de las apacibles comarcas del chauvinismo era su insistencia en una impostada descripción categórica de estímulos y respuestas –remedo, se diría, de la jerga fisicalista o conductista, o acaso del lenguaje ordinario. Dicho de otro modo, el ataque funcionalista contra el fisicalismo y el conductismo daña también los residuos fisicalistas y conductistas que perviven en el seno del propio funcionalismo, en su predilección por ciertos formalismos de descripción de estímulos y respuestas. Como acierta a señalar Block:

> […] functionalists pointed out that physicalism is false because a single mental state can be realized by an indefinitely large variety of physical states that have no necessary and sufficient physical characterization. But if this functionalist point against physicalism is right, *the same point applies to inputs and outputs* […]. That is, on any sense of "physical" in which the functionalist criticism of physicalism is correct, *there will be no physical characterization that applies to all and only mental systems' inputs and outputs*. (Block 1978: 92)

[264] Entiéndase: bien puede ser el caso de que todo sistema capaz de albergar una determinada propiedad funcional sea de hecho un sistema físico; en ese particular sentido, cabría entonces afirmar con verdad que dicha propiedad funcional es una propiedad física –y así lo reconoce Block (1980a: 36). Aun en tales circunstancias, no obstante, seguiría siendo legítimo preguntarnos qué tienen en común todos y solamente los sistemas físicos que albergan dicha propiedad en virtud de lo cual la albergan, y la respuesta a esa pregunta, si el funcionalismo está en lo cierto, no podría darse en lenguaje estrictamente fisicalista.

Es de rigor recordar, por lo demás, que el funcionalismo no se pronuncia acerca de si las circunstancias descritas se darán. Como se ha examinado en detalle, el funcionalismo es para Block una tesis ontológicamente inerme, que no pugna en esa lid ni con el monismo –fisicalista, idealista o neutral– ni con el dualismo o el pluralismo: su contienda no incumbe a qué sean cada uno de nuestros estados mentales, sino a cuáles de sus propiedades los conviertan en estados mentales, y en estados mentales de un determinado tipo.

La cuestión es, a los ojos de Block, que no hay alternativas prometedoras a la descripción física de estímulos y respuestas: una descripción abiertamente mentalista vulneraría el desiderátum funcionalista de que la caracterización de los estados mentales se despliegue sin recurso a vocabulario mentalista, y una descripción funcional nos arrastraría a un escandaloso libertinaje en la atribución de estados mentales. Pero la lección del funcionalismo que parece escapar al escrutinio de Block es que una descripción funcional bien puede ser precisamente la reconstrucción en vocabulario neutral de la caracterización mentalista de nuestras percepciones y acciones, y que la dadivosidad a que nos conduzca a la hora de conceder o denegar el estatus de sujeto de estados mentales bien pudiera ser, al fin y al cabo, una justa generosidad.

En cualquier caso, que el equilibrio entre abstracción funcionalista y concreción fisicalista se vea alborotado por la caracterización funcional de estímulos y respuestas no debería resultar inesperado. Todo decreto que afecte a la descripción que adoptemos de estímulos y respuestas afectará también, *velis nolis*, a nuestra forma de describir estados mentales, y al contrario. Bien podría delinearse así el núcleo de las objeciones levantadas contra Skinner (1957) por Chomsky (1959): la decisión de desterrar de la caracterización de los procesos de aprendizaje toda evocación de lo mental no hizo sino desplazar dichas evocaciones al vocabulario empleado para describir estímulos y respuestas, donde se refugiaban clandestinamente bajo el velo de metáforas y vaguedades de diversa índole. Desde este punto de vista, por otra parte, se hace más estridente la propensión del cognitivismo a dejar agostarse este impulso chomskiano primigenio, replicando sin cuestionarlos los conceptos de estímulo y respuesta tal como se heredaron del conductismo.

Desde esta perspectiva, entonces, el parentesco entre funcionalismo y conductismo –salvo en su vertiente skinneriana– tendría su núcleo no ya en ciertos consensos metodológicos, sino justamente en lo que atañe a la caracterización de estímulos y respuestas. Así, la traducción disposicional de un estado mental bosquejada por el conductismo detallaría las conductas que, encontrándose en tal estado, tendería a emitir el organismo en determinadas circunstancias –el hecho mismo de atravesar el estado mental en cuestión se identificaba entonces con esas tendencias. El funcionalismo, huyendo de las notorias dificultades que acechaban por doquier al análisis disposicional, habría añadido a este esquema la referencia a otros estados mentales –delimitados por caracterizaciones funcionales análogas– que se contaran ya entre las circunstancias en las que el organismo tiende a emitir determinadas respuestas, ya entre las reacciones que el organismo tienda a mostrar en determinadas circunstancias. Pero la caracterización de estímulos y respuestas habría permanecido intacta[265].

[265] No en vano, ni que las divergencias entre el funcionalismo y el conductismo parecieran en su día abismales, ni que la controversia llegara a ser tan encendida –o tan seca, dada la escasa propensión de los conductistas a dar respuesta a las invectivas funcionalistas–, impide a Block (1978) concluir que, en cierto sentido, "[…] functionalism can be seen as a new incarnation of behaviorism" (Block 1978: 63), si

La adopción de un vocabulario extensional, objetivista, para describir estímulos y respuestas es también recogida por Rivière (1991b: 139), en efecto, como uno de los "[…] rasgos de continuidad" –otro es el mecanicismo– que ligan a cognitivismo y conductismo: el cognitivismo habría dado por buena "la premisa conductista" en virtud de la cual "[…] *los enunciados observacionales de la psicología como ciencia deben ser extensionales y formulados en tercera persona del singular*, es decir, la exigencia del objetivismo de método". Sin embargo –cabría argumentar– lo que el funcionalismo nos ha enseñado es cómo podemos reconstruir incluso la atribución de estados mentales a un sujeto en términos extensionales y de tercera persona. Toda vez que esos estados mentales son estados funcionales, podemos identificarlos por el patrón de relaciones que establecen con estímulos, respuestas y otros estados internos, y, al identificarlos así, la intencionalidad y la subjetividad propias de dichos estados mentales quedan aisladas en el seno del referente de las variables sobre las que cuantifica existencialmente el enunciado de Ramsey de la teoría que ampara nuestra atribución de estados mentales. El objetivismo de método podría salvaguardarse, entonces, incluso permitiendo incorporar a la teorización cognitiva descripciones funcionalistas de estímulos y respuestas, siempre que éstas mantuvieran ese carácter extensional y objetivo, y que encontráramos el modo de conjurar la amenaza de circularidad, o de desvinculación del mundo, que preocupa a Block[266], así como el riesgo de incurrir en una desorbitada generosidad a la hora de atribuir vida mental a propios y ajenos.

La convicción de que hemos de disponer de vocabularios diferenciados para la descripción de estímulos y respuestas y la de los fenómenos psicológicos –o, mejor dicho, la constatación de que así sucedía de hecho– está en las raíces del conductismo. Como ha sabido apreciar Yela (1980/1996), la rutina en la redacción de informes experimentales, en la estela de Lloyd Morgan o Thorndike, se había vuelto ya en el paso del siglo notablemente artificiosa:

> El psicólogo va limitándose progresivamente a *describir* los *estímulos* que constituyen la situación del animal –cajas experimentales y laberintos–, las *respuestas* motoras del organismo y las *asociaciones* regulares entre unos y otros hechos […]. A esta descripción se

bien, teniendo como tiene por injustificada la afirmación de que el formalismo de Ramsey permita al funcionalismo identificar nuestros estados mentales en un lenguaje no mentalista, rechaza que esa vigencia del legado conductista en el seno del funcionalismo pueda llevarse más allá. Se aparta Block, por tanto, del dictamen de Leahey (2005: 396), así como el más matizado de Thagard (1992) o el más drástico de O'Donohue y Kitchener (1999: *xx*), todos ellos abordados *supra*. Conviene, en cualquier caso, hacer notar que los dados están cargados: tanto la adopción de un vocabulario fisicalista como la de un vocabulario funcional para la caracterización de estímulos y respuestas resultarían en una prueba de continuidad entre el cognitivismo y el conductismo –por motivos que son ya familiares: basta pensar, en el primer caso en el conductismo de Watson, y en el segundo en el de Skinner.

[266] Sea como sea, no parece difícil argumentar que la decisión de qué clase de vocabulario se emplee para la descripción de estímulos y respuestas –o, como lo plantea Rivière, de qué tipos de enunciados se admitan como parte legítima del lenguaje observacional de la psicología científica– tiene implicaciones que, contra lo que sugiere la expresión "objetivismo de método" empleada por Rivière, desbordan lo estrictamente metodológico (*cf. infra*).

> añade finalmente un análisis de la experiencia subjetiva del animal, de las sensaciones y afectos que en su conciencia acontecen, para *explicar* psicológicamente la conducta observada. (Yela 1980/1996: 168)

Es razonable que Watson execrara esta "*repetición de lo mismo*, en lenguaje mental" que sólo respondía a una "pura inercia de escuela" (Yela 1980/1996: 168). Lo que quedaría, una vez erradicados esos residuos mentalistas, es "[…] la conducta observable que correspondía a aquella conciencia elementalista y sensista, justamente rechazada; es decir, la conducta como movimiento físico […]" (Yela 1980/1996: 168). A la hora de la verdad, sin embargo, Watson llamaría "respuesta a un movimiento muscular, pero también a 'dar una conferencia' o 'construir un rascacielos'" (Yela 1980/1996: 170, *cf.* también Yela 1974) –que tales descripciones presuponían procesos psicológicos es a duras penas cuestionable. De hecho, ya el influyente y enciclopédico examen de *Modern Learning Theory* (Estes *et al*., eds. 1954) había imputado esos cargos a todas las escuelas conductistas. Como de costumbre, Yela (1980/1996) lo compendia con inmejorable precisión:

> Todos [los conductistas] declaran explícitamente que sus conceptos teoréticos o sistemáticos y sus leyes se refieren a los estímulos y respuestas en tanto que "observables físicos", –energías y movimientos– mientras que, por el contrario, sus datos observados se refieren casi siempre a estímulos y respuestas "globales", es decir, a las "situaciones y objetos" a los que el animal responde –barras, laberintos– y a las "acciones" con las que responde –doblar a la izquierda, llegar a la meta, apretar la barra–. Ahora bien, esos objetos y acciones sólo son designables e identificables por su sentido psicobiológico, pero no por su variable contenido físico. (Yela 1980/1996: 174)

La reprimenda es contundente. Sus ecos, que serían medulares en el análisis de Yela (1974) de la noción de conducta, resonaban a lo largo del "*opus magnificentissimum*" –como lo bautizaría Boring (1959)– de Koch (1959a, 1959b, 1959c): ambiciones explicativas prematuras habrían llevado a simplificar en demasía el *explandandum*. A los conductistas, en pocas palabras,

> […] se les recuerda que una psicología de la conducta, estrictamente objetiva, exige considerar la conducta precisamente *qua* conducta, es decir, como algo, desde luego, físico, pero sólo identificable por su significación psicológica, como acción biológica o personalmente significativa con la que el ser vivo responde a una situación definible por lo que para él, o para su adaptación, biológica o personalmente significa […] (Yela 1980/1996: 174)

"Conducta", así pues, comienza por emplearse para designar reacciones corporales observables y físicamente identificables, y luego, subrepticiamente, por medio de aquellas "tretas invisibles" que denunciaran Miller, Galanter y Pribram (1960: 233, *supra*), pasa a referirse a acciones que sólo cabe reconocer por sus rasgos psicológicos o psicobiológicos –como debe ser, salvo en que no cabe eludir, como el conductista pretende, el escrutinio de dichos rasgos al amparo de un vacuo prurito metodológico. También al primer paso de esa maniobra furtiva –el uso de

"conducta" para referirse a reacciones estrictamente corporales– le aguardaba una severa censura. Si bien –qué duda cabe– no hubo de ser el único –él mismo alude como fuente de inspiración a un temprano trabajo de Hamlyn (1953)–, von Wright (1971: 193) hace notar con sorna particularmente sutil que:

> […] only people who have had their talk perverted by behaviorist jargon would think it natural to call such reactions [as salivation, or the jerking of a knee] "behavior" of a dog or a man. (But one could call them behavior of certain glands or of a knee).

No debe admitirse, en todo caso, que la idea de pensar sobre estímulos y respuestas en términos funcionales sea ajena al conductismo, no al menos en su encarnación skinneriana. Antes al contrario, ya en su tesis doctoral sobre *El concepto de reflejo en la descripción de la conducta*, el propio Skinner (1931: 448-449 *apud* Gondra 1992: 57-58) enfatizaba que:

> […] la descripción de la conducta, si ha de ser científica […], tiene que describir el evento no sólo en sí mismo sino en su relación con otros eventos. […] La descripción plena de un evento incluye la descripción de sus relaciones funcionales con los eventos antecedentes.

Un buen ejemplo de cómo Skinner trataría de acatar su desiderátum lo proporcionan los experimentos sobre la fuerza del reflejo que emprendió como parte de la propia investigación doctoral. Así –como bien anota Gondra (1992: 41-42)– en Skinner (1932) "[…] la tasa de presión de la barra fue descrita en función del 'drive', concebido en términos operativos relacionados con la historia de privación o saciedad del animal". Resulta patente que nada pintaba la cualidad placentera del reforzamiento en el esquema explicativo desarrollado por Skinner: los recelos de Watson hacia la formulación que Thorndike había dado a su ley del efecto[267], en cuyas referencias a la satisfacción o el malestar el joven conductista se maliciaba vestigios mentalistas, parecían quedar, así pues, limpiamente salvados[268]. La

[267] A saber:

> *Of several responses made to the same situation those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections to the situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.* (Thorndike 1991: 244, cursiva en el original)

[268] Muy a su pesar, ya Watson acabaría recayendo en veladas alusiones a las nociones mentalistas de las que renegaba –como, según denunciara Chomsky, le ocurriría también a Skinner. Así, por ejemplo, lo señala Kimble (1998: *xi*), quien aduce a modo de prueba de cargo –*cf.* Yela 1980/1996: 170, *infra*, para otras más concluyentes– unas líneas de *Behaviorism*:

> Individual X is lying in bed trying to get to sleep. The arc light on the street shines through a crack in the shade. He wriggles around a bit. It still strikes his eye. […] He puts his head under the covers. There the stuffiness and the heat soon make him put his head out. Again the insistent light strikes him. Then he gets up and does the one sensible thing –he pins a heavy piece of paper over

naturaleza de los conceptos de estímulo y respuesta era una de las vetas de reflexión metateórica en las que Skinner habría de perseverar con mayor tesón. Pronto entendió claramente que –siguiendo una vez más a Gondra (1992), que glosa las palabras de Skinner (1935); *cf.* Skinner (1935: 476-477, *supra*)–:

> […] dado que los estímulos y las respuestas eran muy diferentes en los distintos ensayos de condicionamiento, los términos tenían que designar clases de eventos y no hechos singulares. Los reflejos tenían que ser definidos en términos de clases de movimientos que exhibían unas relaciones regulares, las cuales podían ser observadas en las curvas suaves de los registros acumulativos. Aunque los movimientos eran diferentes en las distintas ocasiones, la relación funcional entre la respuesta y las variables independientes seguía siendo la misma. Ello significaba que los miembros de una clase podían substituirse unos a otros en las secuencias que llevaban a un mismo estímulo reforzante. (Gondra 1992: 49)

Cuál fuera el vocabulario adecuado para la descripción del estímulo se convertiría en una cuestión candente en la controversia entre conductistas y cognitivistas sobre todo a partir, precisamente, de las inclementes críticas que Chomsky (1959) esgrimiera contra las malas prácticas explicativas que a su entender había adoptado Skinner (1957) en su vano intento de armar un análisis conductual convincente de la conducta verbal. Apenas un año después, la polémica encontraba ya eco en las reflexiones sobre la planificación de la conducta desplegadas por Miller, Galanter y Pribram: los argumentos de Chomsky (1959) –escribían Miller, Galanter y Pribram (1960: 33, *supra*)– dejaban a la psicología conductista irremisiblemente abocada a un dilema –recordemos– "entre la ambigüedad y la irrelevancia" del que sólo cabía escapar abandonando la noción de arco reflejo y redefiniendo los conceptos de estímulo y respuesta en el contexto de unidades conductuales de evaluación y operación con bucles retroactivos –las célebres unidades Test – Operate – Test – Exit (TOTE). Cuando el diagrama del bucle de una unidad TOTE se utiliza para describir un mero reflejo, está representando de forma simultánea los tres niveles de abstracción que Miller, Galanter y Pribram proponen diferenciar: flujo de energía, bajo la forma de impulsos neurales, flujo de información y flujo de control. No suele suceder así, sin embargo, en la esquematización de procesos psicológicos más complejos, donde distintos diagramas pueden ser necesarios para describir el flujo de la señal nerviosa, el de la información y el de la transferencia de control. Ahora bien, es el segundo nivel de abstracción, en el que el diagrama representa la transmisión de información, el que de acuerdo con el diagnóstico de Miller, Galanter y Pribram ha

---

> the break in the shade. This response to A leads him into a new environment which no longer contains A as a stimulus. […] The individual gets rid of the stimulus! […] The term "adjustment" […] is convenient […] provided we mean by it the momentary point where the individual by his action has quieted a stimulus or has gotten out of its range. Let us mean by "adjustment", then, something similar to the end of a trial in learning where the animal gets food, sex or water, or becomes oriented away from a stimulus that produces negative reaction, etc. (Watson 1924/1930: 161-162)

sido de uso común en psicología, incluso bajo el mandato del conductismo. En este nivel de abstracción,

> […] no nos interesan las estructuras particulares o formas de energía que intervienen en la producción de la correlación, sino sólo el hecho de que los acontecimientos que tienen lugar en ambos extremos de la flecha están correlacionados. A los psicólogos les resulta absolutamente familiar esta situación, puesto que es a eso a lo que se refieren cuando dibujan en sus diagramas E-R una flecha que va del estímulo a la respuesta, o cuando definen un reflejo como una correlación entre E y R, pero sin decir nada acerca de la base neurológica de esa correlación. (Miller, Galanter y Pribram 1960: 38)

Pese a que Miller, Galanter y Pribram aconsejarían no ya el segundo nivel de abstracción sino el tercero –la transferencia de control– para el estudio de no pocas actividades complejas, es patente que la afinidad entre la práctica conductista de caracterización de estímulos y respuestas y el embrionario programa cognitivista que se está gestando en su trabajo es mucho mayor de lo que suele pensarse. Ellos mismos admitirían de buen grado que:

> El concepto de reforzamiento representa un importante paso hacia adelante en el camino que va desde los arcos reflejos hasta el bucle de retroacción, pero es necesario dar pasos más atrevidos si es que la teoría de la conducta ha de avanzar más allá de la descripción de simples experimentos de condicionamiento. (Miller, Galanter y Pribram 1960: 41)

Ahora bien, una vez envueltos en el seno de un bucle de retroacción, estímulos y respuestas dejaban de aparecer como acontecimientos sucesivos. Tras aducir a modo de sinopsis de sus propias ideas unas palabras de John Dewey (1896) –de quien Watson había sido alumno en la Universidad de Chicago, pero a quien se jactaba de no entender–, Miller, Galanter y Pribram (1960: 40) concluían que "[…p]uesto que el estímulo y la respuesta son correlativos y simultáneos, no hay que pensar que los procesos estimulares precedan a la respuesta, sino más bien que la guían hasta llegar a eliminar con éxito la divergencia". La noción de simultaneidad en la definición de aferencias, estados internos y eferencias, que sería explotada a fondo por Putnam (1967a, 1967b), comenzaba así a cobrar relieve en el cuestionamiento de la concepción conductista del estímulo y la respuesta emprendido por Chomsky (1959) y Miller, Galanter y Pribram (1960)[269].

[269] De hecho, la lectura de Dewey que se invoca en *Planes y Estructura de la Conducta* anticipa también la recuperación del plano teleológico del funcionalismo a la que ha conducido el paulatino cuestionamiento del solipsismo metodológico, las semánticas funcionales y el internismo, y que ha sido recalcada *inter alia* por Bechtel (1988: 185, *supra*). En efecto, Miller, Galanter y Pribram (1960: 53) anotan con indisimulado entusiasmo la observación de Dewey según la cual "[…] estímulo y respuesta no son distinciones de existencia, sino distinciones teleológicas, esto es, distinciones de función, del papel desempeñado, respecto a la acción de alcanzar o mantener un fin" (Dewey 1896: 365). La ironía reside en que, así tomada, pocos reparos habría tenido Skinner en suscribir esta afirmación.

Es posible, no obstante, que el ímpetu que cobra en el funcionalismo el repudio de la aplicación a estímulos y respuestas de definiciones funcionales parejas a las urgidas para los estados mentales tenga su origen también en el carácter primigenio, casi de mito fundacional, que ha adquirido la confrontación con Skinner. Parte de lo que Chomsky (1959) supo ver es que se compadecía mal con la tajante impugnación del vocabulario mentalista que siempre mantuvo Skinner la tenacidad con que él mismo rechazaba la caracterización física de estímulos y respuestas, convencido de que sólo habían de contar como estímulos, en un análisis funcional, aquellos acontecimientos ambientales cuya correlación con las respuestas del organismo hubiera quedado adecuadamente observada y descrita, y de que habían de quedar descritos precisamente en virtud de esas correlaciones –o viceversa, si se trataba de caracterizar la respuesta. Se compadecían mal porque aunque las nociones de estímulo y respuesta quedaran de ese modo razonablemente bien definidas en el contexto de los paradigmas experimentales de aprendizaje, donde el papel de otros estados o procesos psicológicos del organismo había quedado neutralizado o minimizado por el experimentador, al extraerlas de esos contextos quedaban rebajadas a un uso figurativo, en el que la inevitable influencia de esos otros estados o procesos psicológicos simplemente se aglutinaba de forma más o menos intuitiva en la propia noción de estímulo o de respuesta. Éste es el flanco de la ofensiva desatada por Geach (1957) y Chisholm (1957) contra Ryle (1949) –el problema del círculo de lo mental– que, como hemos visto, persiste en el intento de Chomsky (1959) de desarbolar las propuestas de Skinner (1957).

Pues bien, el rechazo de la caracterización funcional de estímulos y respuestas pudo de esta manera parecer consustancial a la empresa cognitivista. Pero las caracterizaciones funcionales de Skinner eran problemáticas a la luz de los argumentos de Chomsky sólo en la medida en que excluían la incorporación sistemática de términos referidos a estados internos, forzando así su incorporación asistemática, subrepticia y, a menudo, despreocupada, bajo un uso figurativo de las nociones de estímulo y respuesta. Si esta deficiencia del proyecto de Skinner queda, como pretende el cognitivismo, subsanada merced a la definición funcional de los conceptos mentalistas, que autoriza el levantamiento de la prohibición que el conductismo había hecho pesar sobre ellos, entonces las únicas razones perdurables para rechazar la tesis skinneriana de que estímulos y respuestas deban venir definidos en términos de sus relaciones recíprocas son el temor a la circularidad y a un desmedido liberalismo en la atribución de estados mentales que Armstrong (1968: 84, *supra*) y Block (1978: 90, *infra*), respectivamente, han expresado con acierto. Si tales temores resultaran infundados, si encontráramos motivos para desbaratarlos, sólo un residual antagonismo con Skinner podría forzarnos ya a insistir en la caracterización física de estímulos y respuestas.

El caso es que incluso Fodor (1991: 30) parece asumir tácitamente una definición funcional de la categoría de respuesta cuando, insistiendo en su argumento de que el vocabulario teórico de la psicología es imprescindible para dar cuenta de lo que tienen en común las distintas encarnaciones físicas de un tipo de

estado mental, anota que sería inútil aducir que esas distintas encarnaciones –en conjunción con las condiciones que completan las cláusulas *cæteris paribus* de las leyes en las que figura el estado mental, *cf. supra*– tienen en común precisamente el hecho de que causan una determinada conducta ya que "[…] by assumption […]" la descripción de esa conducta "[…] too expresses a state-type that no (projectible) microlevel predicate picks out". Desde este punto de vista, así pues, la caracterización funcional de respuestas –presumiblemente también de estímulos, pues el argumento puede replicarse *mutatis mutandis*– es de hecho parte de la estrategia funcionalista de defensa de la autonomía de la explicación psicológica.

De cualquier manera, el funcionalismo conceptual y el funcionalismo empírico –o, si se quiere, Lewis y Fodor– encaran su decidida caracterización categórica de estímulos y respuestas provistos de herramientas bien distintas. Como con perspicacia constatara Block (1978: 68), en la medida en que el proyecto analítico consiste en reconstruir el significado de los términos con los que cotidianamente nos referimos a estados mentales mediante caracterizaciones funcionales que se nutran de las generalizaciones de orden psicológico aportadas por el sentido común, la coherencia con dicho proyecto obliga a adoptar para la especificación de estímulos y respuestas descripciones que puedan hacerse corresponder con las de esas generalizaciones ordinarias. O sea: si nuestro único texto es la psicología ingenua y no consentimos en describir funcionalmente estímulos y respuestas, entonces el único camino que nos queda en describirlos ingenuamente.

Sin mayores desvelos, Block iguala esa descripción inocente de los estímulos con la de "objetos presentes en las proximidades del organismo", y la descripción de las respuestas con la de "movimientos de partes del cuerpo" (Block 1978: 68). Pero resulta evidente que muchas de las generalizaciones que el sentido común explota para la explicación y predicción de la conducta ajena –también, en ocasiones, de la propia– se sirven de descripciones de la conducta más semejantes a lo que Armstrong (1968: 64, *supra*) denominaría "conducta propiamente dicha". Lo mismo es cierto, indudablemente, respecto de los estímulos. Sin embargo, tales descripciones han sido expresamente proscritas por el funcionalismo analítico, bajo la certera sospecha de que escondan referencias a la relación de la conducta –o el estímulo– con estados mentales. De modo que el conjunto de generalizaciones psicológicas del que ha de nutrirse el funcionalismo analítico, si ha de cumplir el doble requisito de ser *vox populi* y de basarse en descripciones físicas groseras de estímulos y respuestas, quedará impredeciblemente menguado –desde luego, bastante más menguado de lo que supone Armstrong, e incluso de lo que supone el propio Block.

Es más: no sería raro que un buen número de las descripciones ingenuas de estímulos y respuestas se revelaran sensibles al análisis funcionalista, y resultaran, finalmente, identificar nuestras conductas, por ejemplo, según los tipos de estados

mentales que suelen provocar –o que suelen provocarlas[270]. Dado que las expresiones con las que habitualmente nos referimos a estímulos, conductas y estados mentales forman campos semánticos tan íntimamente entrelazados, lo extraño sería que pudiéramos mantener a estímulos y respuestas al margen de una relectura funcionalista que –supongamos– iluminara el sentido de nuestras afirmaciones sobre estados mentales. De hecho, parece que nos acercaríamos más al espíritu de las explicaciones psicológicas de sentido común si abandonáramos la jerga esotérica de los estímulos, las respuestas y los estados internos, y comenzáramos a hablar más bien de las cosas que hacemos en la vida, de sus motivos –que suelen ser convicciones, estados de ánimo, deseos, etc., enraizados a su vez en situaciones complejas, cuyo núcleo casi siempre son comportamientos ajenos– y de sus consecuencias –a menudo, entre otras, la modificación de las circunstancias que las motivaron. Dicho de otro modo: si asumiéramos que la categoría "estímulo" y la categoría "respuesta", en animales sociales como nosotros, tienden a solaparse y a infiltrarse la una en la otra. La exclusión de la caracterización funcional de estímulos y respuestas, si todo esto es correcto, constituiría una prescripción –en realidad, una proscripción– que vulnera el talante presuntamente descriptivo del funcionalismo conceptual. Para el funcionalismo empírico, en cambio, permanecen abiertas algunas vías que el funcionalismo analítico ha cegado para sí. Así, Block –que daba por sentado que la madurez de la teorización psicológica nos depararía un vocabulario internista, desprovisto de toda referencia al mundo que no viniese tamizada por su representación interna en el sujeto– se dar por contento con destacar el hecho de que "[…p]sychofunctionalists, on the other hand, have the option to specify inputs and ouputs in terms of internal parameters, e.g. signals in input and output neurons" (Block 1978: 68).

Es precisamente en sopesar las consecuencias de dicha opción –y, en general, de cualquier decisión respecto a dónde tracemos la linde entre la mente y el mundo que habita– en lo que hemos de demorarnos enseguida. Quizá no esté de más insistir

[270] La descripción funcionalista de los estados mentales –lo hemos visto– da parte de razón a la observación –polémicamente atribuida a Wittgenstein (1953), por ejemplo por Budd (1989: 19), pero desde luego presente en Ryle (1949) y en Malcolm (1984, *supra*)– según la cual las conductas vinculadas a un estado mental no son (meros) efectos suyos, sino parte del concepto de ese estado mental. Lo que se apuntaría ahora es que plegar la descripción de la conducta a los mismos dispositivos de caracterización funcional que acordemos para los estados internos haría justicia a la intuición recíproca: también, *viceversa*, los estados mentales vinculados a una conducta son parte del concepto de esa conducta. El reto al que se enfrentaría entonces el funcionalismo consistiría en mostrar que ni la imbricación de los estados mentales en el concepto de la conducta que suscitan ni la de las conductas en el concepto del estado mental que las engendra impiden que el estímulo o el estado mental sean causas y el estado mental o la conducta efectos: no al menos bajo alguna reconstrucción convincente de la noción de causalidad.

Una cierta lectura de esa intuición parecería subyacer al razonamiento de Séneca cuando, tratando de esquivar una objeción a su tesis de que la ira siempre es originada por una ofensa, comenta: "Verdad es que nos irritamos contra los que han de ofendernos; pero nos ofenden con sus mismos pensamientos, y el que medita una ofensa ya la ha comenzado" (*Sobre la ira*, I, III; 2007: 14) –si bien la reflexión de Séneca, como es obvio, se restringe al contexto moral en que nace.

antes en que la trayectoria toda del funcionalismo y el cognitivismo parece a fin de cuentas traslucir en aquella admonición que tanto Tolman (1959, *supra*) como Guthrie (1959, *supra*) acogerían de buen grado pero que sería demoledora para Skinner cuando Chomsky (1959) la pusiera por obra en su crítica de *Verbal Behavior*: podemos incorporar la representación de su entorno que se fragua en los estados internos del organismo a la explicación de su comportamiento dotándonos de las herramientas teóricas precisas, o podemos seguir rehuyendo ese paso y dejar que las propiedades de esos estados internos infecten sordamente un vocabulario conductista cada vez más laxo. Ahí –se diría– palpita la preocupación inicial por el significado que hemos visto germinar en Miller, Galanter y Pribram (1960: 17, *supra*), y que luego evocaría Bruner (1990, 1997), el ímpetu de los modelos computacionales y del solipsismo, la búsqueda de un fundamento naturalizado para la noción de intencionalidad y, a la par, de un ámbito inalienable de eficacia causal de lo mental y de relevancia explicativa de la teorización psicológica.

El mundo en la mente y viceversa

La cuestión de cómo caracterizar estímulos y respuestas se ha entrecruzado a veces con la controversia acerca de la delimitación del contenido semántico de los estados internos –dónde trazar esa linde entre el mundo y la mente que el lugar común emplaza en la piel[271]. En los orígenes del cognitivismo –más precisamente, quizá, en los orígenes del funcionalismo empírico– parece en efecto anidar la convicción de que el modo de especificar los estímulos y las respuestas que hubieran de concurrir en los criterios de una taxonomía madura de los estados mentales sólo podría hacer referencia a sus características *proximales* –" [...] psychological-state tokens were to be assigned to psychological state-types *solely* by reference to their causal relations to proximal stimuli ('inputs'), to proximal responses ('outputs'), and to one another", decía Fodor (1985: 15, *supra*)–, que es tanto como ligar la descripción de estímulos y respuestas a las propiedades físicas de las señales que inciden sobre nuestros órganos de los sentidos, o a las de nuestros movimientos corporales. Pero si la táctica de adoptar una definición funcional de los tipos de estados mentales sobre los que hayan de operar las generalizaciones que darían cuerpo a la psicología científica se

[271] *Cf.* ya Skinner (1974: 17) sobre cómo el conductismo no niega, a su entender, "[…] los hechos que se dan en el mundo privado debajo de la piel"; también, *inter alia*, Block (1986: *passim*), Devitt (1989: 388, *infra*), Place (1999: 380, *supra*).

Entre paréntesis: parece hablarse aquí de la piel como si ésta envolviera sólo el sistema nervioso. Incluso Descartes (1649) –para quien el alma y el cuerpo parecen estar tan lejos entre sí como puedan estarlo sin quedar declaradamente separadas– incluye entre *Las pasiones del alma* los apetitos sensuales, a la vez que explica las emociones como pasiones cuyo origen radica en movimientos no fortuitos de los espíritus animales en el interior de los nervios, los cuales a su vez proceden del alma misma –mientras que en los apetitos proceden del propio cuerpo. Pero si el lindero entre la mente y el mundo fuera la piel, el hambre o la sed se asemejarían más a sentimientos que a sensaciones, pues nada en ellas nos llevaría a traspasar esa extraña aduana.

acompaña de la hipótesis de que el contenido semántico de dichos estados mentales, su carga intencional, depende precisamente del patrón de relaciones funcionales en el que se trenzan –culminando así, de la mano de Schiffer (1986: 134, *supra*), el tránsito de una teoría funcional a una concepción funcionalista de la mente–, lo que nos encontraremos es que la decisión de caracterizar estímulos y respuestas en términos de sus propiedades físicas nos aboca a una posición internista respecto del sentido en que pueda afirmarse que el contenido semántico de los estados mentales desempeña un papel en la explicación psicológica –esto es, nos aboca a la tesis de que no son en realidad las propiedades de las cosas, sino las de la representación que el organismo se forma de ellas, las que están involucradas en dicha explicación.

Ya Dennett (1982), remedando una jerga que evoca intensamente a Quine (1960) y a la tradición conductista, había desbrozado el camino en sentido contrario, delineando con nitidez el compromiso que una posición internista –el solipsismo metodológico por el que Fodor (1980a) abogaba decididamente– parece compelernos a trabar con una descripción estrictamente proximal de estímulos y respuestas:

> Our methodological solipsism dictates that we ignore the environment in which the organism resides –or has resided–, but we can still locate a boundary between the organism and its environment and determine the input and output surfaces of its nervous system. At these peripheries there are sensory *transducers* and motor *effectors*. The transducers respond to patterns of physical energy impinging on them by producing syntactic objects –"signals"– with certain properties. The effectors at the other end respond to other syntactic objects –"commands"– by producing muscle flexions of certain sorts. (Dennett 1982: 25)

Las propiedades semánticas restringidas –o sea, purgadas de toda referencia a hechos mundanos que no se encuentren de algún modo reflejados en el estado del organismo– sobre la que operan las representaciones internas del sistema sólo podrían ser, entonces, aquellas que dependen exclusivamente de la relación entre determinados estados internos de éste y la actividad de transductores y efectores. Donde esa atribución teórica de propiedades semánticas restringidas no alcance, el dictamen es rotundo:

> […W]e should […] treat the transducers as "oracles" whose sources of information are hidden (and whose *obiter dicta* are hence uninterpreted by us), and treat the effectors as obedient producers of unknown effects. (Dennett 1982: 26)

Obiter dicta, que no *rationes decidenci*: la imagen de una semántica, en el fondo, enteramente inerte, cuyo eco en la metateoría cognitiva hemos de someter a un escrutinio más minucioso, queda ya nítidamente perfilada. Lo que adivinamos en esa imagen, en efecto, no es otra cosa que cierta idea del cerebro como una máquina sintáctica, insensible al significado de los símbolos que manipula, que Dennett –al igual que Field (1978) o Fodor (1980a), aunque *cf. infra*– ha defendido con celo:

> This might seem to be a bizarre limitation of viewpoint to adopt, but it has its rationale: it is the brain's eye view of the mind, and it is the brain, in the end, that does all the work […]. Brains are *syntactic engines*, so in the end and in principle the control functions of a human nervous system must be explicable at this level or remain forever mysterious. (Dennett 1982: 26)

Así que por insólita que resulte la idea de tratar la actividad de transductores y efectores como si de oráculos se tratara, no se vislumbran –piensa Dennett– caminos mucho más prometedores –no al menos mientras no pudiéramos construir una noción de significado a la que cupiera atribuir eficacia causal propia, algo que Dennett considera, según parece, fuera de lugar:

> The alternative is to hold –most improbably– that *content* or *meaning* or *semantic value* could be independent, detectable causal properties of events in the nervous system. (Dennett 1982: 26)

Que nuestras creencias y deseos se muestran con sabida tozudez refractarios a lo que ocurra en nuestro entorno, y bajo esa refractariedad determinan el curso de nuestras acciones –uno de los principales núcleos de motivos para adherirse a una concepción internista del significado que recogería Guttenplan (1994b)– queda también nítidamente reflejado en la distinción entre contenido restringido y contenido amplio. En un trabajo titulado, significativamente, "Advertisement for a Semantics for Psychology", Block (1986) trataba de detallar, como quien publicara un anuncio de oferta de empleo, lo que una teoría del significado había de aportar para poder hacerse cargo de la labor de dar cuenta a un tiempo de la insoslayable trabazón con el mundo de nuestras creencias y deseos, y de su no menos palpable escisión. Con ese propósito se detenía a diferenciar dos modos de abordar la taxonomía de entidades significativas –oraciones, proposiciones, etc.–: atendiendo a su extensión o ciñiéndonos únicamente a su intensión. Cuando el referente está involucrado en la tarea taxonómica empleamos, de acuerdo con la propuesta de Block, un criterio de individuación *amplio*; si nos confinamos a considerar el sentido, estamos poniendo en práctica un criterio de individuación restringido, o *estrecho*:

> Wide individuation groups token sentences together if they attribute the same properties to the same individuals, whereas narrow individuation groups sentence-tokens together if they attribute the same properties using the same descriptions of individuals –irrespective of whether the individuals referred to are the same. In other words, narrow individuation abstracts from the question of (i.e., ignores) whether the same individuals are involved and depends instead on how the individuals are referred to […]. (Block 1986: 85)

Si podemos, efectivamente, discernir estas dos formas de clasificar cualquier conjunto de entidades significativas, es trivial, a ojos de Block, que podemos también distinguir dos facetas de su significado, que sería natural denominar significado restringido –o estrecho– y significado amplio, y que se diferenciarían, crucialmente,

en su acatamiento del principio de superveniencia respecto de los estados físicos en que la entidad significativa en cuestión se encarne –la fisiología cerebral, digamos, en el caso de una creencia o un deseo:

> One can think of narrow and wide individuation as specifying different aspects of meaning, narrow and wide meaning. […] Narrow meaning is "in the head," in the sense of this phrase in which it indicates supervenience on physical constitution […], and narrow meaning captures the semantic aspect of what is common to utterances of (e.g.) […] [sentences containing indexicals] by different people. Wide meaning, by contrast, depends on what individual outside the head are referred to, so wide meaning is not "in the head." […] (Block 1986: 85)

A distintas facetas del contenido semántico de un estado mental corresponderían también, así pues, distintas vertientes de las legítimas ambiciones explicativas de la psicología:

> Narrow meaning/content and wide meaning/content are relevant to psychological explanation in quite different ways. For one thing, the narrow meaning of a sentence believed is more informative about the mental state of the believer. Thus narrow meaning (and narrow content) is better suited to predicting and explaining what someone decides or does, so long as information about the external world is ignored. […]
>
> […] But wide meaning may be more useful for predicting in one respect: to the extent that there are nomological relations between the world and what people think and do, wide meaning will allow predicting what they think and do without information about how they see things. (Block 1986: 86)[272]

Así vista, la distinción entre significado restringido y significado amplio parece pensada para hacer justicia al papel de la noción de significado en una de las intuiciones fundacionales del cognitivismo –a saber, que lo que determina la conducta no son tanto los hechos acaecidos en el entorno del organismo sino cómo

[272] Pese a la aparente pulcritud con que las nociones de contenido amplio y contenido restringido parecen deslindar las distintas facetas del significado de creencias o deseos, así como las distintas vertientes de su imbricación en la explicación psicológica, es importante advertir al menos un rasgo de asimetría entre ambas, que el propio Block señala al recordarnos que el significado restringido posee además "[…] another kind of theoretical import: it determines a function from the expressions and contexts of utterance onto referents and truth values […]" (Block 1986: 86). Parece implícito en el hecho de que Block subraye la importancia teórica de esto que, a su juicio, el significado amplio –el referente, el valor de verdad– no determina en cambio una función de expresión y contexto de proferencia a significado estrecho, de lo que cabría deducir que la función a la que alude Block tiene más argumentos de los que se enumeran. En efecto, hemos visto ya que hay diferencias en significado amplio que no tienen repercursión en el significado restringido: "[…] differences in wide meaning that do not involve differences in narrow meaning […] do not cause behavioral differences"[…] (Block 1986: 133).

Por otra parte, vale también la pena advertir, en este mismo sentido, que "[…] despite the misleading terminology, wide meaning does not *include* narrow meaning" (Block 1986: 86).

éste los perciba o conciba, su significado *para* el organismo[273]– sin faltar por ello a lo que seguramente sea el rasgo medular de la propia noción de significado –que aquello que es significativo lo es porque abriga unos ciertos lazos con las cosas del mundo.

> There are two basic facts on which the narrow/wide distinction is based. One is that how you represent something that you refer to can affect your psychological states and behavior. […] The second basic fact is that there is more to semantics than what is "in the head." […] *Since the truth value of a sentence is determined by the totality of semantic facts, plus the relevant facts about the world, there is more to the totality of semantic facts about the sentence than in the speaker's head. The "extra" semantic facts are about what the referring terms in the sentence refer to*. (Block 1986: 88)

Que se entiende como inherente a lo mental una cierta clausura a esos hechos acerca del mundo se aprecia nítidamente en el modo en que Block (1986), apelando una vez más a las diferencias y similitudes entre mi conducta y la de mi *Doppelgänger* en este mundo o el mundo gemelo imaginado por Putnam (1975a)[274], aborda la cuestión de la menor relevancia que de cara a la explicación de la conducta pueda revestir la semántica de los estados mentales entendida bajo criterios externistas –su significado amplio–, por oposición a su semántica estrictamente interna –su significado estrecho:

> Why is narrow meaning relevant to the explanation of behavior, and why is it relevant to the same way for me and my twin? Taking the second question first: since my twin and I are physically identical, all of our representations have exactly the same internal causal roles, and hence the same narrow meanings. But why is narrow meaning relevant to the explanation of behavior in the first place? To have an internal representation with a certain narrow meaning is to have a representation with certain likely inferential antecedents and consequents. Hence, to ascribe a narrow meaning is to ascribe a syndrome of causes and effects, including, in some cases, behavioral effects […]. The reason my twin and I both jump is that we have representations with conceptual roles that have, as part of their syndrome of effects, jumping behavior. The reason that wide

[273] *Cf*. Miller, Galanter y Pribram (1960: 17, 23, *supra*). El papel que esta intuición –decisiva, como hemos visto, en el descrédito del conductismo– desempeñará en la controversia sobre el papel de la semántica de las representaciones internas en la explicación psicológica, a la luz de la tesis de que su eficacia causal ha de plegarse a la de la sintaxis de tales representaciones, será replanteado más adelante.

[274] La historia es ya bien conocida:

> […] we shall suppose that somewhere in the galaxy there is a planet we shall call Twin Earth. Twin Earth is very much like Earth; in fact, people on Twin Earth even speak English. In fact, apart from the differences we shall specify […], the reader may suppose that Twin Earth is exactly like Earth. He may even suppose that he has a *Doppelgänger* –an identical copy– on Twin Earth […].
>
> One of the peculiarities of Twin Earth is that the liquid called "water" is not H_2O but a different liquid whose chemical formula is very long and complicated. I shall abbreviate this chemical formula simply as XYZ. I shall suppose that XYZ is indistinguishable from water at normal temperatures and pressures. In particular, it tastes like water and it quenches thirst like water. Also, I shall suppose that the oceans and lakes and seas of Twin Earth contain XYZ and not water, that it rains XYZ on Twin Earth and not water, etc. (Putnam 1975a: 223)

> meaning is not as relevant to the explanation of behavior as is narrow meaning is that differences in wide meaning that do not involve differences in narrow meaning (e.g., the difference between me and my twin) do not cause behavioral differences. (Block 1986: 133)

O, como tiempo ya tiempo atrás había apuntado el propio Block (1978: 69):

> Let X be a newly created cell-for-cell duplicate of you (which, of course, is functionally equivalent to you). Perhaps you remember being bar-mitzvahed. But X does not remember being bar-mitzvahed, since X never was bar-mitzvahed. Indeed, something can be functionally equivalent to you but fail to know what you know, or [verb] what you [verb] for a wide variety of "success" verbs.

Como se ha dicho, la posibilidad de concebir un *Doppelgänger* ingénito de cualquier organismo, disgregando así sus propiedades funcionales de su historia, viene acompañada a renglón seguido, en el razonamiento de Block (1978), de la apelación a los argumentos de Putnam (1975a) acerca de la Tierra Gemela. Ambos ejercicios de razonamiento modal confluyen según Block en la conclusión internista de que:

> […i]f functionalism is to be defended, it must be construed as applying only to a subset of mental states, those "narrow" mental states such that truth conditions for their application are in some sense "within the person." (Block 1978: 69)

Con esto, lo que la psicología reclama de una teoría del significado puede en definitiva, de acuerdo con el diagnóstico de Block (1986), condensarse en tres arduas tareas:

> First, a semantics for psychology should have to have something to say about what the distinction between narrow and wide meaning comes to and, ideally, should give accounts of what the two aspects of meaning are. Second, the theory ought to say why it is that narrow and wide meanings are distinctively relevant to the explanation and prediction of psychological facts (including behavior). Third, the theory ought to give an account of narrow meaning that explains how it is that it determines a function from the context of utterance to reference and truth value. (Block 1986: 87)[275]

También McDermott (1986) ha sostenido que la atribución de creencias y deseos con contenido restringido es por fuerza atribución de creencias y deseos *de re* acerca de estímulos y conductas descritos en términos proximales, es decir, atribuciones cuya verdad depende únicamente de hechos objetivos acerca ya de la estimulación recibida por los órganos sensoriales, ya de los movimientos musculares del organismo[276]. De hecho, la argumentación de McDermott (1986) está dedicada a

[275] Cf. Frege (1892, *infra*).

[276] Estrictamente, el vocabulario de la atribución de actitudes proposicionales con contenido restringido incluiría según McDermott, aparte de la descripción proximal de estímulos y conductas, partículas lógicas y un deíctico anafórico empleado para referirse al propio organismo.

descartar la viabilidad de otras formas de caracterización del contenido restringido que no sean la apelación a propiedades proximales de estímulos y respuestas a la que nos constreñía Dennett (1982): a saber, que una actitud proposicional con contenido restringido sea una actitud acerca del entorno del organismo, o que lo sea acerca de propiedades fenomenológicas de los objetos de dicho entorno.

En ambos casos, se trata de argumentos modales que siguen los surcos del trabajo de Putnam (1975a) sobre la Tierra Gemela. Imaginemos, en primer lugar, que uno cree que el agua es líquida, pero su *Doppelgänger* en la Tierra Gemela –donde todo, salvo la composición química del agua, es exactamente igual que en nuestra Tierra– no cree tal cosa: no puede creerla, ya que sus creencias se refieren al compuesto químico XYZ, que es completamente indistinguible del agua pero no es agua –el agua, suponemos, es H_2O. Por tanto, la simple creencia de que el agua es líquida –una actitud proposicional acerca del entorno del organismo– no puede ser una creencia con contenido restringido, dado que no cabe atribuirla con inmutable valor de verdad a dos organismos idénticos[277], o, lo que es lo mismo, dado que su semántica –cuando menos, su valor de verdad– depende de hechos mundanos.

Más laborioso, en cambio, resulta desechar la tesis, que se atribuye a un trabajo inédito de Fodor titulado "Narrow Content and Meaning Holism", de que el vocabulario que describe el contenido restringido esté formado por términos referidos a propiedades fenomenológicas. Fodor habría tratado de sustituir "*S* cree que el agua es líquida" por "*S* cree que existe algo transparente, inodoro, insípido, etc., que es líquido". El escenario contrafáctico que McDermott blande contra Fodor es bastante más alambicado:

> Suppose that on Twin-Earth everything is green, including ripe tomatoes (I mean Twin-tomatoes). However, Twin-sun does not produce light, but a strange sort of radiation which, when it falls on the green tomatoes, causes them to emit *light* of wavelength r. So Twin-tomatoes *look* red. They are Twin-red. But they are not red, they are green –they have the kind of surface which, when illuminated by white light, reflects green. (McDermott 1986: 127-128)

Si todo esto es así, de nuevo, podemos imaginar que uno cree que hay un tomate rojo frente a sí, pero su *Döppelganger* no cree tal cosa: no puede creerla, ya que sus creencias se refieren al color rojo gemelo (efecto de la radiación del Sol Gemelo sobre una superficie verde), el cual es indistinguible del color rojo, pero no es rojo. Por

---

Por otra parte, es importante señalar que tanto la atribución de actitudes proposicionales con contenido restringido especificado en otros términos, siempre que éstas supervengan en actitudes *de re* acerca de estímulos y conductas proximales, como la atribución de actitudes proposicionales con contenido amplio, son compatibles a juicio de McDermott con la tesis de que una psicología cognitiva basada en la noción de contenido restringido debe ceñirse al vocabulario prescrito.

[277] Lo que se exige –aclara McDermott– no es que la creencia de mi *Doppelgänger* tenga el mismo valor de verdad que la mía, como en Stich (1978), sino que la atribución de dicha creencia tenga el mismo valor de verdad tanto si se refiere a mí como a mi *Döppelganger*. Por esa razón el argumento de Stich excluye también deícticos referidos al propio organismo, mientras que el de McDermott los consiente.

tanto, la creencia de que hay algo rojo (redondeado, etc.) frente a mí –una actitud proposicional acerca de propiedades fenomenológicas de los objetos del entorno– no es una creencia con contenido restringido, dado que no cabe atribuirla correctamente a dos organismos idénticos. Así que, en suma, si la caracterización del contenido restringido de una actitud proposicional no puede hacer alusión ni al entorno ni a sus propiedades fenomenológicas, deberá ceñirse a mencionar estímulos y conductas descritos en términos proximales, *quod erat demonstrandum*.

Dos objeciones a su planteamiento, que considera menores, se detiene a examinar McDermott. La primera es que las actitudes proposicionales acerca de estímulos y conductas especificados en términos proximales son una mera ficción teórica: nadie tiene creencias ni deseos acerca de cosas como la radiación electromagnética que estimula su retina o los impulsos de los nervios motores –la divergencia entre el concepto de representación mental que manejaría la picología cognitiva y el propio de la actitudes proposicionales atribuidas en la vida cotidiana se hace demasiado radical. Pero McDermott despacha expeditivamente la cuestión: si nos parece inaceptable la atribución a un ser humano de creencias sobre, por ejemplo, la radiación que estimula su retina, es porque damos por hecho que negaría creer tal cosa si le preguntáramos. Esto no debe inquietarnos, sin embargo, puesto que sucede siempre que se trata de actitudes *de re*: Edipo deseaba *de re* casarse con su madre, aunque no lo deseara *de dicto* y, por tanto, negase desearlo. A la réplica de que a toda actitud *de re* subyace una actitud *de dicto* –Edipo deseaba, al fin y al cabo, casarse con Yocasta–, requisito que incumplirían las presuntas actitudes sobre estímulos y conductas descritos en términos proximales, opone McDermott diversas actitudes *de re* que no vienen respaldadas por ninguna actitud *de dicto*: sería el caso de las creencias y deseos de otros animales, o de niños que aún no han adquirido el lenguaje, que no podríamos en modo alguno atribuir *de dicto*, así como de aquellas de nuestras creencias y deseos que sobrepasan nuestra propia capacidad de conceptualización, o al menos de verbalización. Un ejemplo de estas últimas sería la creencia de que una tela de un color que no sabemos nombrar es *de ese color*, creencia que explicaría nuestra habilidad para reconocerla entre otras de colores parecidos: es fácil atribuir *de re* dicha creencia ("*S* cree que la tela es del color que realmente es"), pero no habría, según McDermott, modo de atribuirla correctamente *de dicto*.

La segunda objeción a la que McDermott se enfrenta es la de que una psicología como la que él recomienda obvia aspectos de las creencias y deseos de un organismo que pueden constituir diferencias importantes entre ése y otros organismos. Por ejemplo: el sentido común reconoce una evidente diferencia entre una persona que cree que vive en un entorno natural y una persona que cree ser un cerebro en una probeta, diferencia que resultaría inapresable en el vocabulario de una psicología basada en el contenido restringido de actitudes *de re* acerca de estímulos y conductas descritos en términos proximales. A falta de contraargumentos, McDermott reconoce la discrepancia entre su propuesta y la concepción cotidiana de las actitudes proposicionales, y trata de circunscribirla: al menos, no se trata de una renuncia al realismo, dado que la diferencia que queda obviada no es la que media entre vivir en

un entorno natural y ser un cerebro en una probeta, sino apenas entre *creer* una cosa y creer la otra.

Pero la principal debilidad del argumento, que McDermott reconoce, no es sino la premisa de que la verdad de *toda* creencia acerca de *X* depende de *X*, sobre la que descansa a su vez la de que mi *Doppelgänger* no cree que el agua sea líquida, o que el tomate frente a sí es rojo. Si esa premisa es sustituida por la de que la verdad de una creencia acerca de X dependerá de X o de su contrapartida epistémica relevante –así, la verdad de una creencia acerca del agua dependerá según el caso de hechos sobre el agua o de hechos sobre el compuesto que surge en los manantiales de la Tierra Gemela–, el argumento queda sin efecto. Esta táctica –piensa McDermott– supone una revisión de la concepción cotidiana de las actitudes proposicionales y su relación con la verdad; si bien no sería del todo raro –habrá de concederse– que necesitemos revisar nuestra idea intuitiva del contenido de las creencias y los deseos si hemos de adaptarla a escenarios con mundos gemelos al nuestro y soles bajo cuya luz los colores son otros. En todo caso, la conclusión a la que nos conduciría esta estrategia sería reconocer la viabilidad de una psicología cognitiva basada en la atribución de actitudes proposicionales con contenido restringido especificadas como actitudes acerca del entorno del organismo, actitudes que serían semejantes en todo a las cotidianas –creencias y deseos, clasificadas según su contenido amplio– excepto en su relación con la verdad.

Ahora bien, el problema es que una psicología cognitiva de ese tenor carecería según McDermott de potencia explicativa, al no permitir su vocabulario la formulación de leyes psicológicas. La razón es que formular una generalización exige introducir al menos una variable bajo cuantificación universal tanto en el antecedente como en el consecuente de un enunciado condicional ("para todo *a*, si *Fa*, entonces *Ga*"), de modo que al extraer instancias concretas de la generalización obtengamos un enunciado condicional cuyo tema sea común para antecedente y consecuente ("si Fa_1, entonces Ga_1"). Esto es factible en una psicología centrada en el contenido restringido de actitudes *de re* acerca de estímulos y conductas descritos en términos proximales –como la recomendada por McDermott–, y también en una psicología centrada en contenidos amplios, pero no en una psicología centrada en el contenido restringido de actitudes acerca del entorno. En el primer caso, tanto antecedente como consecuente tratarían acerca de estímulos y conductas descritos en términos proximales. En el segundo caso, ambos tratarían acerca de características del entorno, sea en sí mismas o como objeto de actitudes proposicionales del organismo. Pero en el caso de una psicología centrada en el contenido restringido de actitudes acerca del entorno, uno de los términos del enunciado condicional –argumenta McDermott– tendría que versar acerca de estímulos o conductas descritos en términos proximales, y el otro acerca de características del entorno, lo cual bloquea la generalización.

> In short, attitudes to the environment seem just the thing for commonsense wide psychology. For commonsense psychology aims to explain the link between two things which involve the environment […]. But when we factor out the physics […], and then

> factor out the physiology of the retina and the […] muscles, so that the focus of psychology is narrowed down to the gap between input and output, then the environment is no longer relevant. And then it's attitudes to inputs and outputs which are just the thing. (McDermott 1986: 131)

Lo que McDermott obvia es que el núcleo de su argumentación –que– presupone que estímulos (o, *mutatis mutandis*, respuestas) habrán indefectiblemente de venir descritos en términos proximales, físicos. Dicho de otro modo: nada en el razonamiento de McDermott proscribe que algunas generalizaciones de la psicología vinculen estímulos descritos en un vocabulario físico proximal a estados internos con contenido descrito en los mismos términos, mientras que otras vinculen estímulos descritos en términos funcionales a estados internos descritos como actitudes acerca del entorno distal del organismo en virtud de su contenido restringido –o amplio. Algo así, por lo demás, es de hecho lo que parece ocurrir en la práctica de la teorización psicológica, como ha sabido señalado Burge respecto a la descripción proximal –es decir, miográfica– de la conducta[278]:

> But this construal has almost no relevance to psychology as it is actually practiced. […] Much behavior is intentional action; many action specifications are non-individualistic. […]
>
> For example, much "behavioral evidence" in psychology is drawn from what people say or how they answer questions. Subjects' utterances (and the questions asked them) must be taken to be interpreted in order to be of any use in the experiments; and it is often assumed that theories may be checked in experiments carried out in different languages. […] Many attributions of non-verbal behavior are also intentional and non-individualistic, or even relational: she picked up the apple, pointed to the square block, tracked the moving ball, smiled at the familiar face, took the money instead of the risk. (Burge 1986: 44-45)

Al fin y al cabo, el funcionalismo empírico puede en principio hacer suyo *cualquier* modismo que la psicología diera en adoptar para la descripción de estímulos y respuestas, siempre que éste no entorpeciera el trámite de poner cada tipo de estímulo en relación con los (tipos de) estados internos que suele causar, y cada tipo de respuesta en relación con los (tipos de) estados internos de los que suele ser efecto. De hecho, pese a que tanto la vertiente analítica del funcionalismo como su vertiente empírica confluyeran, por distintos caminos, en abominar de la caracterización funcional de estímulos y respuestas, Block no puede dejar de conceder, al menos, que tal línea de trabajo resulta plenamente congruente con los planteamientos funcionalistas acerca de la naturaleza de lo mental, puesto que " […] this version of functionalism treats inputs and outputs just as all versions of functionalism treat internal states" (Block 1978: 90).

Así, oídos sordos de las advertencias de Block o de Armstrong parece hacer, por ejemplo, Pylyshyn (1984) –a quien, por lo demás, no sería del todo verosímil relegar como un heterodoxo. Pese a fijar una estricta exigencia de que las aferencias que

[278] *Cf.* al respecto las penetrantes observaciones de Yela (1980/1996: 170, 174, *supra*).

reciban los transductores sensoriales vengan caracterizadas en términos físicos –en un "vocabulario proyectable", es decir, susceptible de formar parte de una ley física– si es que el mecanismo en cuestión ha de poder catalogarse como transductor (Pylyshyn 1984: 165)[279], y pese a forjar después una esmerada defensa de dicho requisito ante la tentación de concedernos rienda suelta para, a la hora de dar cuenta del funcionamiento de los transductores, generalizar sobre estímulos descritos en un vocabulario cognitivo, o funcional, lo cierto es que Pylyshyn no sólo no tiene ningún reparo en aventurar la formulación de generalizaciones sobre estímulos (o respuestas) caracterizados en términos cognitivos *siempre y cuando* no estemos tratando de explicar el funcionamiento de un transductor sensorial, sino que, de hecho, la ferviente insistencia en que sólo así podremos llegar a entender la conducta humana –a capturar las generalidades relevantes (Pylyshyn 1984: *passim*)– es quizá el corazón de su tesis. Baste recordar en este sentido la perseverancia con que Pylyshyn argumenta que:

> […] it is (a) the environment or the antecedent event *as seen or interpreted by the subject*, rather than as described by physics, that is the systematic determiner of actions; and (b) actions performed with certain *intentions*, rather than behaviors as described by an objective natural science such as physics, that enter into behavioral regularities. (Pylyshyn 1984: 9, *supra*)

La motivación para aferrarse a una descripción fisicalista de estímulos y respuestas que hemos entrevisto en Armstrong (1968) y Block (1978) –el temor a una explicación circular, desasida del mundo objetivo– parece, con todo, atisbarse también en Pylyshyn. No en vano la idea en torno a la que orbita su defensa de que es imperioso para el cognitivismo describir las aferencias de los mecanismos de transducción sensorial en un vocabulario fisicalista es, precisamente, que si no lo hiciéramos así, si cediéramos a la tentación de dejar que las categorías cognitivas impregnen también la explicación de dichos mecanismos, "[…] we [would] stand little chance of systematically relating behavior to properties of the physical world" (Pylyshyn 1984: 168). Esta coincidencia deja que se adivine ya con cierta nitidez una posible conclusión: las restricciones fijadas por Pylyshyn tal vez sean suficientes para aliviar la preocupación que lleva a Armstrong y a Block a fijar unas mucho más rígidas –tal vez lo sean, esto es, si además de aplicarlas a la transducción sensorial las trasladamos al ámbito del control psicomotriz. A fin de cuentas, restringir la inadmisibilidad del vocabulario funcional a estos dos dominios explicativos no

[279] Quiere decirse que la propia función que corresponde desempeñar a los mecanismos de transducción sensorial en el seno de la teorización cognitiva –tal como la concibe Pylyshyn– requiere que sus aferencias vengan descritas en lenguaje fisicalista. Otros dos requisitos del mismo orden son que los transductores operen aislados del resto del sistema cognitivo –son, dice Pylyshyn, *stimulus-bound*, o *data-driven*–, y que formen parte de la arquitectura funcional de dicho sistema –sus funciones son primitivas y se ejecutan sin que medien procesos simbólicos (Pylyshyn 1984: 153-154). Cualquier estructura psicológica que pudiéramos hipotetizar y en la que no se verificasen estos requisitos no sería, por definición, un transductor –no al menos en el sentido estipulado por Pylyshyn.

contravendría el corolario que el propio Block extrae de su razonamiento: que el funcionalismo parece apresar los estados mentales –inmovilizarlos, si se quiere, como un entomólogo sus especímenes– "[…] only at the periphery –i.e., through physical, or at least nonmental, specification of inputs and outputs" (Block 1978: 64, *supra*). De lo que se trataría, antes bien, es de delimitar con más cuidado lo que verdaderamente debemos considerar periférico.

A juicio de Bechtel (1988: 179), si bien los impedimentos que Block achaca tanto al funcionalismo empírico como al conceptual a la hora de caracterizar estímulos y respuestas sin caer en el liberalismo o en el chauvinismo parecen reflejar una insuficiencia real de ambas concepciones de lo mental, cabría al funcionalismo una escapatoria que preserva su potencia explicativa y su observancia de los cánones naturalistas. Decidir qué sistemas –organismos, máquinas, ángeles…– son legítimos sujetos de estados y procesos mentales, con independencia de la caracterización que hayamos adoptado de los estímulos y respuestas de dichos sistemas, exigiría tener en cuenta los propósitos a los que sirven los estados y procesos en cuestión. Es, en suma, el problema de la descripción de estímulos y respuestas, estrechamente entrelazado con el de los límites de la atribución de estados mentales, lo que habría de conducir al funcionalismo hacia una perspectiva teleológica. Merced a esa clave teleológica podríamos, pues, expulsar a los célebres sistemas homunculares pergeñados por Block –ya se sabe: la población china, la economía boliviana– de la república de lo mental, puesto que ninguno de ellos despliegan su actividad "[…] para cumplir con el mismo género de fuerzas de selección que las que actúan sobre mí" –o, en general, sobre un sujeto humano (Bechtel 1988: 182). Podríamos, también, de acuerdo con el mismo expediente, ofrecer o no carta de ciudadanía en esa nación nuestra a distintas variedades de máquinas, en la medida en que "[…] interactúen y se adapten […] íntimamente a las exigencias de su medio" (Bechtel 1988: 184), o no; a juicio de Bechtel (*ibid*.), "[…] este análisis da la respuesta correcta para juzgar la cuestión de si los computadores tienen mentes".

Podría ser. No procede ahora desglosar las virtudes y quebrantos de la propuesta, ni cotejarla con el viejo juego de imitación imaginado por Turing (1950a, *supra*) –y que acabaría por llevar su nombre–, pero acaso sí apuntar que la ruta por la que Bechtel se encamina parece desembocar sin remedio en un proyecto de caracterización funcional de estímulos y respuestas. No es fácil –al menos– vislumbrar respuestas de otra índole a la pregunta que la apuesta teleológica hace inminente: cómo especificar a qué se refieren giros del estilo de "el mismo género de fuerzas de selección que", "un medio del género correcto" (Bechtel 1988: 182), "géneros de presiones de […] selección […] radicalmente diferentes" (Bechtel 1988: 183), etc. Dicho de otro modo: lo que se está fraguando es una caracterización de los estímulos que el organismo capta y de las respuestas que emite en el seno del ambiente que habita en términos de las relaciones de tales estímulos y respuestas con las presiones evolutivas a las que se halla sometido el organismo. En la jerga más abiertamente teleológica –e intencional, y vagamente lamarckiana– que Bechtel prefiere:

> Lo que la perspectiva teleológica nos exige hacer no es simplemente considerar las interacciones causales al identificar funciones, sino considerar cómo esos procesos causales están contribuyendo a las *necesidades* del organismo, tal como están especificadas por exigencias del medio. (Bechtel 1988: 182)

Pero esto es ya una caracterización funcional de estímulos y respuestas, si bien sólo en estado larvario: resulta presumible que estímulos –o respuestas– de características físicas sumamente diferentes puedan suponer para diferentes organismos –de la misma o distintas especies– idénticas presiones evolutivas y, por tanto, que hayan de quedar taxonomizados como del mismo género. Así, por ejemplo, que un estado mental dado sea o no un caso de miedo no habrá de depender de las características físicas del estímulo que lo elicita o la respuesta que induce, sino de hechos como que la configuración estimular indique la presencia de un depredador, que la respuesta pueda entenderse como (un intento de) evitación, sea por huida, mimetismo, o por cualquier otra de las múltiples formas de escabullida con que la naturaleza colma a los moralistas de ejemplos de cobardía, o, en fin, que los otros estados internos que suelan acompañarlo incluyan una temerosa ansiedad anticipatoria, o alguna noción del propio desvalimiento, o lo que se quiera. Al tratar de adelantarse a las objeciones que Block pudiera plantearle en cuanto a la viabilidad de la tarea de detallar qué presiones evolutivas son del mismo género que las que ha soportado *Homo sapiens* y cuáles no, el propio Bechtel toma decidida, aunque tal vez no deliberadamente, la dirección de la caracterización funcional. Así, inspirándose en la distinción de Mayr (1974) entre sistemas cerrados y abiertos, sugiere Bechtel que:

> Los atributos psicológicos parecen sólo ser aplicables a aquellos organismos que adoptan una estrategia abierta y aprenden qué conductas realizar. Tales sistemas tienen que ser sensibles a la información sobre sus medios y ser capaces de procesar esta información para determinar las respuestas apropiadas. Esto sugiere que podríamos ser capaces de desarrollar una explicación general de los procesos mentales en términos de sus papeles en sistemas abiertos (p.ej., como procesos que figuran en el procesamiento de información a partir de un medio que, a continuación, determina estrategias de acción). (Bechtel 1988: 183)

Pero "una explicación general de los procesos mentales en términos de sus papeles en sistemas abiertos" no es más que una explicación funcionalista con la inocua *addenda* de que el sistema cuyos procesos internos se trata de explicar es capaz de aprender. Convertir eso en una explicación teleológica de corte evolucionista exige dar un paso más: caracterizar los estímulos que inciden sobre el sistema y las conductas con que el sistema incide sobre su medio en términos de sus respectivos papeles en una urdimbre de presiones evolutivas, en la que se entrelazan con otros estímulos, estados internos del organismo y conductas, en lugar de hacerlo en términos estrictamente fisicalistas –es decir, suscribir una caracterización funcional de estímulos y respuestas.

Por cierto, todo esto ha de llevarnos a hacer una anotación en los márgenes del atlas de las concepciones funcionalistas de la mente que venimos de trazar: del mismo modo que dotar a la explicación funcional de basamentos teleológicos nos compromete con la extensión del análisis funcionalista al ámbito de los estímulos y las respuestas, el intento de conjugar la perspectiva teleológica con los postulados homucularistas –tal como queda bosquejado en los trabajos de Dennett, Lycan o Bechtel– no puede sino reforzar ese compromiso. La razón es la siguiente: supongamos que, efectivamente, la atribución de estados y procesos mentales a un sistema depende de que las presiones evolutivas a las que está sometido sean, en un sentido por definir, del mismo género que las que han moldeado la adaptación a su entorno de especies como la nuestra; en tal caso, o bien la atribución de estados y procesos mentales a los subsistemas especializados que la teorización psicológica describa como componentes de nuestro sistema cognitivo *no* está justificada porque tales subsistemas no se enfrentan a presiones evolutivas apropiadas –*ergo* el homuncularismo es falso–, o bien *sí* está justificada porque los homúnculos de alguna manera se compadecen de las presiones evolutivas que nos aquejan a nosotros, o porque padecen las suyas propias, en su propio hábitat –padecen, se compadecen: dicho sea sin incurrir en una irónica petición de principios respecto a su capacidad de sufrimiento. A la vista está que el primer itinerario conduce ineluctablemente a la renuncia al homuncularismo. El segundo nos forzaría a pensar en el pandemónium de homúnculos que desde este punto de vista poblarían nuestra mente como una suerte de enjambre de huéspedes simbióticos endosomáticos cuya supervivencia, sencillamente, dependiera de la nuestra. Pero no es cierto, en general, que los simbiontes endosomáticos, ni siquiera los endosimbiontes (es decir, los simbiontes endocelulares) exhiban rasgos adaptados a las presiones evolutivas de sus anfitriones, y, lo que es peor, esta estrategia no nos permite deslindar los subsistemas cognitivos a los que queremos atribuir estados mentales de otros subsistemas orgánicos a los que no queremos hacerlo –los alveolos pulmonares o las glándulas sudoríparas, *cf*. *supra*. Para recorrer, por último, la tercera ruta trazada será imprescindible valernos de caracterizaciones funcionales de estímulos y respuestas. No es imaginable de qué otro modo podrían las aferencias y eferencias de un subsistema cognitivo cualquiera resultar equiparables a los estímulos y conductas del organismo que lo alberga, si no es especificando que se trata de equiparar el papel de aquéllas de cara a la adaptación del homúnculo a su entorno con el papel de éstas de cara a la adaptación del organismo al suyo –incluso la noción de adaptación, probablemente, requiera aquí una lectura funcional, desprovista como quedaría de todo vínculo con las de reproducción y herencia. Además de toparnos con expedicionarios del darwinismo neural, inspirados por los trabajos pioneros de Edelman (1978, 1989), no sería raro que en estos parajes diéramos también con el propio Dennett. Al ser preguntado por la principal cuestión científica o filosófica en que hubiera cambiado de opinión a lo largo de su carrera, Dennett ha escrito:

> I've changed my mind about how to handle the homunculus temptation […].
>
> Notice that computers have been designed to keep needs and job performance almost entirely independent. Down in the hardware, the electric power is doled out evenhandedly and abundantly; no circuit risks starving. At the software level, a benevolent scheduler doles out machine cycles to whatever process has highest priority, and although there may be a bidding mechanism of one sort or another that determines which processes get priority, this is an orderly queue, not a struggle for life. […] Neurons, I have come to believe, are not like this. […B]rain cells –I now think– must compete vigorously in a marketplace. For what?
>
> What could a neuron "want"? The energy and raw materials it needs to thrive –just like its unicellular eukaryote ancestors and more distant cousins, the bacteria and archaea. Neurons are robots; they are certainly not conscious in any rich sense –remember, they are eukaryotic cells, akin to yeast cells or fungi. If individual neurons are conscious then so is athlete's foot. But neurons are, like these mindless but intentional cousins, highly competent agents in a life-or-death struggle, not in the environment between your toes, but in the demanding environment of the brain, where the victories go to those cells that can network more effectively, contribute to more influential trends at the virtual machine levels where large-scale human purposes and urges are discernible. […]
>
> Intelligent control of an animal's behavior is still a computational process, but the neurons are "selfish neurons," […], striving to maximize their intake of the different currencies of reward we have found in the brain. And what do neurons "buy" with their dopamine, their serotonin or oxytocin, etc.? Greater influence in the networks in which they participate. (Dennett 2008)

Así pues, parece que la única estrategia disponible para dotar de credibilidad a una lectura homuncularista del funcionalismo pasa por adoptar una caracterización funcional de estímulos y respuestas. Dicho de otro modo: sólo tal caracterización funcional genera una concepción de lo mental suficientemente liberal como para que una intrincada red de células nerviosas pueda convertirse en un homúnculo.

Asunto bien distinto es –naturalmente– que esta estrategia sea todo lo provechosa que sus adalides creen. Pero con independencia de cuánta verdad impregne la imagen homuncularista de la mente y el cerebro, cabe anotar como establecida la conclusión de que el vocabulario pertinente para la descripción de estímulos y respuestas ha de variar según la naturaleza del proceso psicológico abordado –o incluso del proceso fisiológico en cuestión, si las observaciones de Dennett estuvieran bien encaminadas. Mientras que la investigación experimental sobre procesamiento perceptivo temprano seguramente no pueda pasar sin descripciones estrictamente físicas de los estímulos –en términos, por ejemplo, de la distribución de energía que presentan–, no es aventurado afirmar que ya en la investigación sobre procesos de asignación de recursos atencionales se haga imprescindible recurrir en algún momento a una caracterización parcialmente funcional de la estimulación. Ni que decir tiene que en el ámbito de la investigación sobre aprendizaje, sobre memoria, pensamiento, motivación o emoción, la descripción funcional de los estímulos, y también la de las respuestas, se impone con toda naturalidad; en cuanto a la psicopatología y la psicología anormal, es inaudito que se plantee siquiera la caracterización física de "estresores" o de síntomas. La psicología del lenguaje, más allá del área de la percepción del habla o de los

mecanismos de control articulatorio, difícilmente subsistiría si se la privara de la caracterización de unidades lingüísticas –palabras, oraciones, textos– agrupadas en clases de equivalencia funcional más o menos exacta. El estudio del control locomotor, por último, sin duda requiere de un vocabulario tan netamente físico como el de los procesos perceptivos básicos, esta vez de orden cinemático. En todos los casos que –por así decir– carecen de la salida al mar de los estímulos y las conductas, la descripción de los estímulos o de las conductas relevantes asume una carga de determinación funcionalista de su extensión –un *sentido* funcional, pues. En un modelo exhaustivo y completo del sistema cognitivo –en una idealizada descripción total de la mente, del estilo de las que Putnam (1997: 35, *supra*) tacha de ficticias–, esa carga habría de quedar liquidada –cabe pensar– en la descripción fisicalista de los procesos sensoriomotores –digamos, por perseverar en la metáfora geográfica– colindantes. Pero quizá ni siquiera en esa teoría acabada pudiera cancelarse la intensión funcionalista que impregna la descripción de estímulos y respuestas en la investigación sobre cualquier proceso psicológico que no sea rigurosamente periférico[280]. Quizá, con otras palabras, nuestra concepción de lo mental haya de quedar en alguna medida exenta, ingrávida, separada del mundo por pequeños resquicios que le otorguen la maleabilidad que parece precisar.

Aún si así fuera, no parece que nos aceche en estas regiones la falacia de circularidad de la que Armstrong (1968) y Block (1978) recelaban. Es más, no parece que haya nada reprochable en semejante división del trabajo explicativo, salvo que seamos ciegos a su naturaleza e ignoremos sus tramoyas. Nuevamente resuena aquí la insistencia de Aristóteles en "[…] no buscar del mismo modo el rigor en todas las cuestiones, sino, en cada una según la materia que subyazga a ellas y en un grado apropiado a la particular investigación. (*Ética Nicomáquea* I: 1098*a*, *supra*). De hecho, la idea de estructura de la conducta de Miller, Galanter y Pribram (1960) cuadraba ya con lo esencial de estas observaciones, siempre que las contemplemos desde el lado de la conducta. Nada más citar los influyentes desarrollos del análisis lingüístico de constituyentes presentados por Chomsky (1957), constatan Miller, Galanter y Pribram que:

> Los psicólogos, cuando se han ocupado de la conducta verbal, rara vez han mostrado aversión a inferir la existencia de unidades molares tales como "palabras" o, incluso, "significados", aunque las respuestas de las que la percepción dispone realmente sean meras cadenas de sonidos, representaciones acústicas de los fonemas correspondientes. Desgraciadamente, sin embargo, el psicólogo describe normalmente la conducta –o algún aspecto de la conducta– a un único nivel, y deja que sus colegas utilicen su propio sentido común para inferir lo que ha sucedido en otros niveles. (Miller, Galanter y Pribram 1960: 25)

[280] Suponiendo, dicho sea de paso, que existan procesos rigurosamente periféricos, íntegramente aislados de todo influjo que no sea el de las propiedades físicas de la señal sensorial o motora: es decir, que sea correcta en alguna medida la hipótesis de modularidad de Fodor (1983).

Ahora bien, tampoco la renuncia a órdenes superiores de descripción de estímulos y respuestas sería un empeño sensato. De lo que se trata más bien es –claro está– de desarrollar en toda su complejidad una teoría de la estructura jerárquica de la conducta:

> El registro meticuloso de cada contracción muscular, aunque alguien hiciera acopio de valor bastante para intentar llevarlo a cabo, seguiría siendo insuficiente, ya que no contendría los rasgos estructurales que caracterizan las unidades molares; y aquellos rasgos molares deben ser inferidos sobre la base de una teoría acerca de la conducta. Nuestras teorías de la conducta, en este sentido del término, siempre han sido implícitas e intuitivas. […]
>
> En aquellos casos afortunados en los que se nos ofrecen descripciones adecuadas de la conducta […] resulta completamente obvio que ésta se organiza simultáneamente en varios niveles de complejidad. Nos referiremos a este hecho con la expresión: "organización jerárquica de la conducta". (Miller, Galanter y Pribram 1960: 25)

En suma, así pues, la descripción de estímulos y conductas para los fines explicativos de la psicología es funcional hasta los huesos tanto como lo es la de los estados internos. Para otros fines –no sería difícil fraguar el argumento– la descripción de estímulos y conductas tal vez no haya de ser funcional, pero tampoco sería entonces descripción *de estímulos y conductas* –lo sería acaso de los mismos eventos, pero no en tanto que estímulos y conductas.

Que el desarrollo teórico del funcionalismo, y parejamente el trabajo experimental en las ciencias cognitivas, se encauzara hacia la definición funcional de estímulos y respuestas acarrearía, por último, una rara lección. Aparte de en el ensayo de su propia refutación, el legado fundamental del conductismo habría consistido –según el dictamen de MacKenzie (1977), que Yela (1980/1996) considera cicatero y trata enérgicamente de ensanchar– en "[…] algo así como una fenomenología práctica, que puede servir y está sirviendo de propedéutica al estudio experimental y teorético ulterior" (Yela 1980/1996: 181). Pero debería resultar transparente, a la luz de estas investigaciones, que es precisamente esa fenomenología práctica lo que ha de inspirar el despliegue de la teorización psicológica más allá de las restricciones impuestas por las formulaciones tempranas del funcionalismo y el cognitivismo.

Un cerco invisible

El más diáfano y conciso recuento de lo que ha supuesto en nuestras reflexiones acerca de la mente la idea de que ésta traza un cerco invisible en derredor suyo, del cual el mundo queda fuera, es tal vez el que puede encontrarse en uno de los más tenaces críticos de tal concepción:

> According to individualism about the mind, the mental natures of all of a person's or animal's mental states (and events) are such that there is no necessary or deep individuative relation between the individual's being in states of those kinds and the nature of the individual's physical or social environments.

> This view owes its prominence to Descartes. It was embraced by Locke, Leibniz, and Hume. And it has recently found a home in the phenomenological tradition and in the doctrines of twentieth century behaviorists, functionalists, and mind-brain identity theorists. (Burge 1986: 39-40)

Contra el individualismo –es bien sabido– Burge ha presentado varios argumentos modales en los que una persona, *A*, posee determinada noción –la de aluminio, la de artritis– de forma sesgada o incompleta, y se nos invita a imaginar entornos contrafácticos –mundos posibles– en los que se mantiene idéntico el estado físico del organismo de *A* la vez que se registran cambios profundos en cuanto al referente de sus nociones –el aluminio es un metal ligero, la artritis no ha sido aislada como síndrome o lo ha sido bajo otros parámetros–: la conclusión presumiblemente intuitiva es que A carecería en tales entornos de las nociones que le atribuíamos en el entorno actual, es decir, que se producirían diferencias significativas en sus estados mentales pese a que, *ex hypothesi*, el estado físico de su organismo se mantiene inalterado. Por tanto, según Burge:

> [...M]ental states and events may in principle vary with variations in the environment, even as an individual's physical (functional, phenomenological) history, specified non-intentionally and individualistically, remains constant. (Burge 1986: 41)

Como no podía ser de otro modo, Burge es plenamente consciente de que su argumento refuta la superveniencia de los estados mentales del organismo sobre sus estados físicos y, con ello, la tesis de identidad psicofísica –tal vez también determinadas lecturas del funcionalismo. Ahora bien, esto –matiza el propio Burge– no equivale a una refutación del materialismo[281]: hacerlo equivaler supondría obliterar la distinción entre relaciones de individuación y de composición, así como entre individuación y causación. Dicho de otro modo:

---

[281] Tampoco –desde luego– es el materialismo sobre los estados mentales una consecuencia trivial de la tesis de superveniencia, pues bien podría admitir dicha tesis un insobornable paralelista.

Las diferencias entre individuación, causación y composición que Burge esgrime contra el individualismo habían madurado ya, bajo distintas formas, en el desarrollo teórico del monismo anómalo de Davidson (1970, 1973a, 1974a, *cf.* también 1963 y 1993), que pretende ofrecernos una concepción sólida de la causación mental en la que el reduccionismo sea tan poco pertinente como el dualismo. Pero de la refutación del individualismo ensayada por Burge a la del reduccionismo que Davidson anhela no hay, ni mucho menos, un paso franco. Entre los obstáculos que lo plagan destaca la sospecha de que, dada esta distinción entre causación e individuación, Burge estaría encaminado a un instrumentalismo a la manera de Dennett. Es decir, ¿no se convierte la individuación de los estados, eventos y procesos mentales del organismo en una cuestión de perspectiva o estrategia explicativa, causalmente inerte y, por tanto, prescindible al cabo si se adopta una estrategia explicativa de orden estrictamente físico, o fisiológico –la estrategia de diseño, o *design stance*, en la terminología de Dennett? Trasladada al ámbito de la argumentación de Davidson, esta sospecha ha dado pie al nutrido enjambre de acusaciones de epifenomenalismo que han venido recayendo sobre el monismo anómalo –entre ellas, las de Dretske (1989), Fodor (1989) y, con particular estruendo dialéctico, las de Kim (1978, 1979, 1984b, 1993b, 1993c y, sobre todo, 1989), que hemos examinado *supra*.

> It is simply not a "trivial consequence" of materialism about mental states and events that the determinants of our behavior supervene on the states of our brains. This is because what supervenes on what has at least as much to do with how the relevant entities are individuated as with what they are made of. If a mental event *m* is individuated partly by reference to normal conditions outside a person's body, then, regardless of whether *m* has material composition, *m* might vary even as the body remains the same. (Burge 1986: 46)

En cualquier caso, el carácter local de la causalidad es algo que Burge se muestra de todo punto dispuesto a conceder –no así que de él se derive que nuestra individuación de los fenómenos estudiados deba proceder también según criterios de índole local, pues, se deduce, no es forzoso que nos valgamos en nuestras taxonomías de criterios de individuación exclusivamente fundados sobre relaciones de causalidad. La réplica a esa maniobra –imputación de una explicación por acción a distancia– es adivinada por Burge; la contrarréplica –*equivocatio*– es tajante[282]. Vale la pena evocar en detalle el razonamiento:

> It is plausible that events in the external world causally affect the mental events of a subject only by affecting the subject's bodily surfaces; and that nothing (not excluding mental events) causally affects behavior except by affecting (causing or being a causal antecedent of causes of) local states of the subject's body. One might reason that in the anti-individualistic thought experiments these principles are violated insofar as events in the environment are alleged to differentially "affect" a person's mental events and behavior without differentially "affecting" his or her body: only if mental events (and states) supervene on the individual's body can the causal principles be maintained.
>
> The reasoning is confused. The confusion is abetted by careless use of the term "affect", conflating causation with individuation. Variations in the environment that do not vary the impacts that causally "affect" the subject's body may "affect" the individuation of the information that the subject is receiving, of the intentional processes he or she is undergoing, or of the way the subject is acting. It does not follow that the environment causally affects the subject in any way that circumvents its having effects on the subject's body. (Burge 1986: 47-48)

Antes al contrario –insiste Burge:

> We may agree that a person's mental events and behavior are causally affected by the person's environment only through local causal effects on the person's body. Without the slightest conceptual discomfort we may individuate mental states so as to allow distinct

[282] La imputación según la cual en la supuesta refutación del individualismo se apela a mecanismos de acción a distancia merece una respuesta particularmente expeditiva por parte de Burge:

> Some authors have suggested similarities between […] action-at-a-distance theories in physics, and non-individualistic theories in psychology. The analogies are tenuous. […]. Unlike action-at-a-distance theories, […] [anti-individualism] does not appeal to action at a distance. It is true that aspects of the environment that do not differentially affect […] physical movement […] do differentially affect the explanations and descriptions. This is not, however, because some special causal relation is postulated, but rather because environmental differences affect what kinds of laws obtain, and the way causes and effects are individuated. (Burge 1986: 50)

Uno de los autores veladamente aludidos es sin duda Block (1986: 91).

> states (types or tokens) with indistinguishable chemistries, or even physiologies, for the subject's body. Information from and about the environment is transmitted only through proximal stimulations, but the information is individuated partly by reference to the nature of normal distal stimuli. Causation is local. Individuation may presuppose facts about the specific nature of a subject's environment. (Burge 1986: 48)

Junto con esta objeción, Burge rechaza taxativamente otras dos. En primer lugar, la idea de que las entidades teóricas postuladas por la psicología *deben* sobrevenir en estados, eventos o procesos fisiológicos, donde la noción relevante de *deber* es de orden metodológico, resulta, según Burge, errónea a la luz de lo que resulta habitual en otras ciencias naturales. En segundo lugar, la idea de que la teoría psicológica *debe* recoger las similitudes, más que las diferencias, entre el sujeto actual y el sujeto contrafáctico del argumento modal de Burge, tan pronto como se trata de darle alguna nitidez, viene a reposar a juicio de Burge sobre argumentos ya rechazados: es, en suma, "vaga o entimémica" (Burge 1986: 51). Bien puede verse, en fin, que el propósito de Burge no es otro que desposeer al individualismo de todo apoyo en el materialismo o en los principios del método científico, de suerte que, sin constituir una descripción ajustada de la teorización psicológica y a falta de justificación ontológica, el individualismo se revele como mera "ideología metafísica" (Burge 1986: 49).

Tras el asedio al individualismo, entonces, llega la hora de reclamar tesis positivas:

> Ascription of intentional states and events in psychology constitutes a type of individuation and explanation that carries presuppositions about the specific nature of the person's or animal's surrounding environment. Moreover, states and events are individuated so as to set the terms for specific evaluations of them for truth or other types of success. We can judge directly whether conative states are practically successful and cognitive states are veridical. For example, by characterizing a subject as visually representing an *X*, and specifying whether the visual state appropriately derives from an *X* in the particular case, we can judge whether the subject's state is veridical. Theories of vision, of belief formation, of memory, learning, decision-making, categorization, and perhaps even reasoning all attribute states that are subject to practical and semantical evaluation *by reference to standards partly set by a wider environment*. (Burge 1986: 54)

La respuesta más común a estos argumentos, de acuerdo con Burge, es aceptarlos, acaso resignadamente, para el caso de las atribuciones cotidianas de actitudes proposicionales, pero rechazarlos sin titubeos para el caso de los estados internos asignados a los organismos en la teorización cognitiva: así, "[…n]on-individualistic aspects of mentalistic attribution have been held to be uncongenial with the purposes and requirements of psychological theory" (Burge 1986: 41). Pero esta tesis no descansaría –insiste Burge– sobre la realidad de las explicaciones psicológicas vigentes, sino que supondría un programa revisionista sobre el lenguaje teórico y los presupuestos básicos de la psicología. La única justificación de un programa semejante provendría del argumento según el cual el *explanandum* de la psicología es

la conducta de los organismos, entendiendo el concepto de conducta de forma tal que ésta no se viera afectada por los entornos contrafácticos imaginados por Burge[283]. Este argumento es rebatido por Burge mediante *(i)* el cuestionamiento del concepto de *conducta* que se asume (el cual precisaría la exclusión de acciones intencionales, conducta verbal semánticamente especificada, o, en general, de toda conducta dirigida a objetos, dejando así a la psicología sin apenas *explanandum*), así como, por supuesto, mediante *(ii)* el cuestionamiento de la tesis general de que la psicología sea la ciencia de la conducta y no, por ejemplo, de ciertas capacidades molares del organismo como la memoria, la formación de creencias, la comprensión del habla, o la categorización (Burge 1986: 45).

En una esforzada defensa del carácter solipsista de la explicación psicológica, Jacob (1987) rechaza los argumentos modales de Burge bajo la acusación de que incurren en el error de asignar a un estado mental propiedades pertenecientes a la expresión lingüística que lo describe –el mismo error que conduce según Jacob a la conclusión de que observación está impregnada de teoría. Según Jacob, lo que los escenarios contrafácticos elaborados por Burge describen son comunidades lingüísticas diferentes, y lo que muestran es en todo caso que el significado de las palabras con las que describimos los estados mentales de un sujeto –Bob, en el ejemplo imaginado por el propio Burge– puede depender, en ese sentido, del entorno distal. Ahora bien:

> On Burge's view, the differences between the two linguistic communities enter *ipso facto* the content of Bob's thought.
>
> [...] However further premisses are needed to justify [...] the inclusion of the environment (whatever it contributes to the content of her utterances) in the content of her thoughts. (Jacob 1987: 81)

Esas premisas adicionales están según el diagnóstico de Jacob abocadas a fracasar. La razón es que:

> Including the experts' knowledge of the […] extension of "arthritis" […] in the wide meaning of such words […] is to confuse (first-order) conditions of satisfaction or truth-

[283] El énfasis en la conducta –en cierta concepción de la conducta, más bien– como *explanandum* de la psicología estaba presente, en efecto, en la argumentación de Field (1978: 62), y lo estaría también en Jacob (1987: 88-89), que advierte ya el flanco abierto por Burge y trata de protegerlo:

> […] in arguing in favor of the legitimacy of narrow (or inferential) content, I linked the notion of narrow thought […] to behavior […]. Have I not […] begged the whole question in presupposing a narrow individuation of behavior which is, according to […] [Burge], not to be had? […]
>
> On the face of it, this seems like a serious objection. But I think there is less to the argument than meets the eye. The reason being that the dispute about the proper individuation of intentional mental states can just be run again about the proper individuation of behavior. (Jacob 1987: 88-89)

Cf. también, en la misma línea, Dennett (1982: 10, *infra*).

> conditional facts or states of affairs with (second-order) knowledge (or mastery) of the aforementioned facts or conditions. (Jacob 1987: 85)

Dicha inclusión inhabilitaría, además, a la teoría semántica derivada de los planteamientos de Burge para dar cuenta de la posibilidad del error –que, como vimos, se configura como una de las marcas distintivas de la intencionalidad ya desde Brentano y Frege, si no desde Platón. Aquí es donde da su fruto la distinción, de aire lockeano, entre representaciones mentales privadas, de primer orden, y sus expresiones lingüísticas, públicas, de segundo orden: dar cuenta del error requeriría, de acuerdo con Jacob, adjudicar a las primeras un contenido restringido, individualista, y sólo a las segundas un contenido amplio, que incluya factores sociales:

> Remember that in the actual community Bob informs his doctor of his belief that he has arthritis in his thigh. By Burge's own admission, the natural thing to say then is that Bob has an inaccurate (or deficient) concept of arthritis. But if the concept mentally represented in Bob's mind were identical to the (wide) meaning of the word "arthritis" (as medical experts use it), then it would be impossible for Bob to have a deficient concept. While the wide meaning of the word is constituent of the semantic value of an utterance which is a public representation, the concept is a constituent of a thought which is a mental representation. (Jacob 1987: 85)

En definitiva, Jacob redescribe el escenario contrafáctico imaginado por Burge como uno en el que el protagonista expresa verbalmente una proposición distinta, en la medida en que son distintas sus condiciones de verdad, de la que expresa en el escenario fáctico, pero dado que él no es consciente de tales diferencias, éstas no permean su pensamiento, que permanece definido por su contenido restringido. Aunque si un habitante del escenario fáctico y otro del escenario contrafáctico describieran con las mismas palabras las creencias del protagonista sobre la artritis estarían atribuyéndole creencias con diferente contenido amplio –es decir, contenido con su plena carga semántica, extendida hasta abarcar las propiedades de aquello a lo que se refiere la creencia– el caso es que las creencias del protagonista no habrían variado[284].

Lo cierto es que el esfuerzo por dar cuenta de la posibilidad de albergar creencias erróneas se ha ido perfilando como uno de los caballos de batalla de la controversia acerca de si los estados mentales de un sujeto pueden quedar

[284] Esta redescripción deja entrever, por otra parte, que el acecho de cierta suerte de explicación por acción a distancia en el planteamiento de Burge –*cf. supra*– no pasa inadvertido a Jacob:

> However one way to justify the anti-individualist inference from public language to intentional mental representation might be to claim, as Burge (1982 […:] 107) does, that "propositional attitude attributions which put the terms in oblique occurrence will thus affect the content of the propositional attitudes". But this would amount to the view that ascriptions do not reveal, that they create thought-contents, and would thus be inconsistent with the realist presuppositions embodied in our common ascription practice. (Jacob 1987: 91)

adecuadamente caracterizados sin mención del entorno en el que éste se desenvuelve. Imaginemos –como nos invita a hacer Guttenplan (1994b: 289)– que tras mirar por la ventana digo algo así como "Ese pájaro es un verderón", pero que lo que vi en realidad no fue más que un temblor del follaje, que tomé erróneamente por un pájaro. La cuestión es, entonces, si mis palabras expresan una creencia sobre un pájaro inexistente –o, suponiendo que hubiera dicho sencillamente "Hay un pájaro ahí", una creencia falsa–, o si más bien no albergo creencia alguna sobre ningún pájaro –pues no hay tal pájaro–, ni por tanto la expresan mis palabras, por mucho que yo *crea* albergar tal creencia: "[…o]n this strong externalist stance" –apunta Guttenplan (1994b: 289)– "[…] propositional attitudes become opaque to their possessors. We can think we believe and desire various things –that our attitudes have certain contents– though we might well just be wrong". La intuición de que conocemos los contenidos de nuestra propia mente, de que incluso gozamos de alguna suerte de autoridad cuando los describimos, pesaría entonces –como bien señala Guttenplan– en favor de una concepción estrictamente interna de dichos contenidos, pues aunque "[…] I may be wrong about there being a bird, […] how can I be wrong about my believing that there is one?" (Guttenplan 1989b: 291). Acaso con más fuerza que una intuición que, después de todo, bien podría estar equivocada –lo cual, en la medida que la intuición atañe a cierta imposibilidad de equivocarnos, no dejaría de ser un signo casi cruel de la menesterosidad de nuestro juicio–, conviene tener en cuenta que la interpretación externista de mis creencias acerca del pájaro inexistente conduce a un *regressus ad infinitum* que Guttenplan no parece advertir: si no es posible que yo albergue una creencia sobre un pájaro que no existe, tampoco puede serlo que la albergue sobre una creencia igualmente inexistente –es decir, que crea erróneamente que creo que el pájaro es un verderón–, de modo que tendríamos que decir que en realidad sólo creo creer que creo que el pájaro es un verderón –el pájaro, por cierto, que creí ver, o que creí creer que veía… Parece, en suma, que tanto los errores en el uso de conceptos –errores categoriales como el del paciente de artritis– como los errores perceptivos –como el de ver un verderón donde sólo había unas hojas– resultan poco afines a la idea externista del contenido de las creencias y reclaman una caracterización menos abierta[285].

Como se ha adelantado, otro de los pilares sobre los que Jacob erige su defensa de la necesidad de preservar modismos restringidos de descripción de nuestros

[285] En el recuento de las virtudes de la concepción computacional de la mental con que Horst (1996) encabeza su incisivo relato de sus defectos, cobra especial relieve la aparente facilidad con que ésta puede hacerse cargo, desde los presupuestos internistas que venimos discutiendo, no ya sólo de la posibilidad de errores como los mencionados, sino, además, de "[…] several traditional pitfalls associated with the hard cases presented by illusions, hallucinations […], and other deviant cases of perception and cognition" A ello agrega Horst, también, que la noción de computación habría permitido al cognitivismo eludir a un tiempo "[…] the Meinongian tendency to postulate nonexistent entities and the opposite inclination to identify the contents of intentional states with the extramental objects towards which they are directed" (Horst 1996: 44) –cabría, desde luego, discutir que esas dos propensiones sean opuestas.

estados psicológicos, de los que el entorno quede expulsado, es una cierta concepción de la conducta en tanto que *explanandum* de la psicología –una concepción igualmente restringida, y que Burge (1986: 45, *supra*) cuestiona. Las intuiciones que respaldan al internismo, sin embargo, son también aquí vigorosas, y quedan a la luz en el sencillo ejemplo elegido por Guttenplan:

> Suppose that I reach for my binoculars just after insisting that I saw the bird in the tree. The obvious explanation for my action would seem to mention, among other things, my belief that there is such a bird. However, since if the externalist is right, then I just do not have any such belief, it is unclear how to explain my reaching for the binoculars. (Guttenplan 1994b: 290)

Es más: resulta a duras penas discutible que existe algún sentido razonable de la noción de conducta bajo el cual mis acciones cuando busco los prismáticos *habiendo visto un pájaro* y cuando los busco *creyendo haberlo visto* cuentan como las mismas acciones, pero la propuesta de Burge vetaría esa lectura al obligarnos a asumir que unas y otras vienen causadas por creencias irremediablemente diferentes –si es que logramos siquiera reconstruir el sentido en que creer que uno ha visto un pájaro pudiera, bajo esos parámetros, llegar siquiera a constituir una creencia[286].

El propósito último de Jacob, en cualquier caso, es mostrar que aquello en lo que el entorno distal pueda contribuir a determinar el contenido de las representaciones mentales no contradice, *pace* Burge, la tesis de superveniencia psicofísica respecto de la eficacia causal de lo mental. Lo que se trata de preservar, en suma, es un principio general según el cual

> […] the relevance of the environment to the individuation of some of an individual's mental intentional states is no threat to the supervenience of the mental causes of the individual's behavior upon his or her brain states. (Jacob 1987: 79)

El caso es que, como hemos visto, que la eficacia causal de los estados mentales haya de guardar una estricta superveniencia con los estados físicos en los que aquellos se encarnen no es algo que Burge parezca por la labor de discutir: sus reparos incumben más bien a la idea de que nuestros recursos de individualización y taxonomización de dichos estados mentales deban doblegarse inexorablemente a los pliegues de esa eficacia causal. A pesar de que en ella se han puesto en juego concepciones en apariencia tan distantes de la explicación psicólogica, de la conducta o de la

[286] Esta es la idea que aparentemente subyace al tercer conjunto de motivos para adherirse a una concepción internista de los estados mentales que enumera Guttenplan (1994b), y que él plantea en términos de la posibilidad de rendir cuentas de las diferencias y semejanzas entre mi conducta cuando habiendo visto un pájaro busco los prismáticos y la de un *Doppelgänger* mío que hiciera exactamente lo mismo, pero en cuyo entorno no hubiera tal pájaro, o incluso la mía propia si, transportado instántaneamente y sin mi conocimiento a dicho entorno, me aprestara a buscar los prismáticos. Se diría que la carga de razonamiento modal –y el esfuerzo de la imaginación– que requiere el planteamiento de Guttenplan es innecesario para asentar la cuestión, pero nada, desde luego, depende de que así sea.

naturaleza de las creencias y del conocimiento de nuestra propia vida mental, cierto sustrato compartido parece aflorar después de todo[287].

Por mucho que las observaciones de Jacob puedan debilitar la crítica del individualismo armada por Jacob, o que la perspectiva de que podamos construir la semántica que Block reclama para la psicología pueda mitigar sus efectos, lo cierto es que los argumentos de Burge dejan abierta una brecha entre la forma en que cotidianamente pensamos acerca de nuestras propias creencias y deseos y la forma en que se articulan las representaciones internas que, de acuerdo con una concepción computacional de la mente, subyacen a esas creencias y deseos. No son sólo ya los saberes o las creencias que una comunidad comparte y que se cristalizan en el lenguaje: Cummins (1989), por ejemplo, señala certeramente –al hilo de un argumento de Stich (1983: 165), que también ciertas trayectorias históricas impregnan lo que entendemos coloquialmente por una creencia o un deseo:

---

[287] Curiosamente, toda la controversia entre Burge y Jacob parece descansar sobre diferentes particiones de un mismo espacio conceptual. Mientras para Burge la tesis individualista es que *ningún* estado mental precisa para su individuación de referencia alguna al entorno distal (físico o social) del organismo (*cf.* Burge 1986: 39-40, *supra*) –es decir, el individualismo es una tesis excluyente–, Jacob adjudica tal carácter excluyente al antiindividualismo –para el individualismo, en cambio, "[…] facts from an individual's environment are sometimes relevant to the individuation of the individual's thought. Sometimes they are not" (Jacob 1987: 79). En suma, lo que Burge (1986) defiende bajo el nombre de antiindividualismo y lo que Jacob (1987) defiende bajo el nombre de individualismo parecen por momentos ser una y la misma doctrina.

> Burge takes his and Putnam's [1975a] thought-experiment to show that the environment (social and non-social) is inevitably part of the content of an individual's thoughts or beliefs. While I take them to suggest that there are two aspects to the concept of belief or perhaps there are two belief concepts and this is the reason why they elicit different intuitions. […]
>
> But on the anti-individualist view, only one set of intuitions makes sense […]. Only one aspect of the belief concept would be respectable –the broad or wide (as opposed to the narrow) aspect. There is therefore an important argumentative asymmetry between the anti-individualist and his opponent here. (Jacob 1987: 86-87)

Esa asimetría argumentativa deposita el peso de la prueba sobre los hombros del externista –"[…] because he is an eliminativist with respect to the narrow (or functional) mode of individuation", dice Jacob (1987: 88). La cuestión es que el externista, según parece, podría perfilar la asimetría justo en sentido contrario: pese a sus ecuménicas declaraciones de principios y a las acusaciones cruzadas, los argumentos de Jacob, no menos que, *mutatis mutandis*, los de Burge, tienden después a perfilar sus propias conclusiones en términos excluyentes. Así, el propio Jacob (1987: 87-88) continúa su argumentación asignando la noción de contenido restringido, individualista o funcional, a la psicología científica, y la noción de contenido amplio, no-individualista o relacional, al intento cotidiano de obtener información sobre el entorno a partir de las creencias de los demás, o bien a particularísimas investigaciones sociológicas sobre el nivel de conocimientos de una población respecto de un asunto determinado –por ejemplo, la naturaleza fisiopatológica de la artritis. Entretanto, Burge (1986: 53-64) aplica argumentos antiindividualistas incluso a las representaciones postuladas en la teoría del procesamiento visual temprano de Marr (1982), con lo que hace difícil ver en qué vertientes del funcionamiento de la mente pudiera tener cabida el contenido restringido, máxime a la vista de una enumeración de ámbitos no-individualistas de teorización psicológica que incluye "theories of vision, of belief formation, of memory, learning, decision-making, categorization, and perhaps even reasoning" (Burge 1986: 54, *supra*).

> […] beliefs *are* individuated in a way that is sensitive to history. […T]he belief my duplicate expresses with the words "I sold my car for a thousand dollars" is false because he didn't own the car in question, whereas the belief I express with the same words is true. We therefore have different beliefs, though by hypothesis we are computationally (indeed physically) equivalent. The problem is that my duplicate never acquired title to the car in question; I did. Hence, history matters to belief contents. If you are interested in belief contents, then you will do well to formulate an account that is sensitive to historical properties. (Cummins 1989: 84)

Pero si nos trasladamos de la vida cotidiana a la teorización psicológica –mejor dicho, a la teorización psicológica tal como ésta se modula bajo la idea de que los procesos cognitivos son procesos computacionales–, las cosas resultan muy otras. Tanto, de hecho, que parece imponerse la conclusión de que las representaciones internas sobre las que operan esos procesos computacionales no pueden identificarse ya con las creencias o los deseos que saturan nuestro día a día, o, ni siquiera, con el sustrato de su semántica:

> It doesn't follow from this, of course, that *representation* is sensitive to historical properties. In fact, nearly the opposite follows: Since data structures aren't sensitive to historical properties, it follows that belief aren't data structures. Moreover, it follows that beliefs don't inherit their contents from constituent data structures, as the R[epresentional] T[heory of] I[ntentionality] claims. (Cummins 1989: 84)[288]

Esto, sin embargo, no hace –a ojos de Cummins– más que poner de manifiesto que los propósitos explicativos de la psicología cognitiva no coinciden plenamente con los de la psicología coloquial –como, por otra parte, podría decirse de cualquier empresa científica al cotejarla con aquellos aspectos de nuestra praxis cotidiana que pudiéramos consignar como su germen.

> Data structures are insensitive to all sorts of things –such as historical properties– to which beliefs (and the other propositional attitudes) are exquisitely tuned. There is a good reason for this. As Stich points out, the C[omputational] T[heory of] C[ognition] doesn't want to explain why my duplicate can't sell my car. Or, to put it as Ned Block does […], some differences in belief are not legitimate sources of psychological variance. (Cummins 1989: 84)

---

[288] O, si se quiere, más tajantemente –en una formulación que anticipa la controversia sobre el papel de la semántica en la explicación psicológica, en el que se tratará de profundizar *infra*–:

> It is not at all obvious that cognitive functions must be, or even *can* be, intentionally specified. The C[omputational] T[heory of] C[ognition] […] seeks an *individualistic* psychology, i.e., a psychology that focuses on cognitive capacities of the kind that might be brought to bear on radically different environments. If the anti-individualistic position with regard to intentionality is right […], then the explananda of and individualistic psychology cannot be specified intentionally. It follows that the C[omputational] T[heory of] C[ognition] shouldn't –indeed *mustn't*– concern itself with intentionally specified explananda. What the anti-individualist arguments of Putnam and Burge prove from the viewpoint of the C[omputational] T[heory of] C[ognition] is that beliefs and desires aren't psychological states in the sense of 'psychological state' of interest to the C[omputational] T[heory of] C[ognition]. (Cummins 1989: 140)

Ya Pylyshyn (1984: 2, 263, 271, *supra*) –recordemos– había anticipado que son los recursos teóricos de una ciencia los que perfilan el contorno de su campo de estudio, con lo que el inventario de fenómenos de los que puede dar cuenta bien puede ir corrigiéndose a tenor del propio avance de la disciplina –en realidad, diría Pylyshyn, no puede establecerse de ninguna otra manera. Pero las divergencias que ahora parecen salir a la luz no atañen a regiones más o menos remotas del *explanandum* de la psicología natural que pudieran resultar inesperadamente ajenas para la psicología cognitiva, o viceversa, sino a una de las reivindicaciones nucleares del funcionalismo: que una teoría computacional de los procesos psicológicos puede aparejarse al patrón explicativo fundamental de la psicología de sentido común –la articulación de creencias y deseos.

El vínculo entre la cuestión de cómo caracterizar el contenido restringido y la divergencia entre el concepto cognitivista de representación mental y el de las actitudes proposicionales es explícito en el trabajo de Baker (1985). La concepción funcionalista de las actitudes –que éstas se definen por sus relaciones con estímulos, conductas y otros estados internos– exige según Baker una interpretación restringida de su contenido. Hay, eso sí, diversos modos de construir tal interpretación, que Baker rechaza considerar equivalentes: la originaria, regida por el criterio con que Putnam (1975a) definía la tesis de solipsismo metodológico, es que la atribución de una actitud proposicional no debe presuponer la existencia de ninguna entidad individual específica que no sea el sujeto al cual se atribuye dicha actitud (Putnam 1975a: 220); Fodor (1980a: 66-67), en cambio, oscilaría entre la idea de que la explicación de la conducta requiere que la atribución de actitudes sea opaca y la de que requiere que no tenga en cuenta la verdad o la referencia a entidades individuales específicas; tanto Block (1978: 69, *supra*) como Stich (1983: 165) o Davidson (1987: 443-444), por último, han contribuido, *inter alia*, a propagar la idea de que el contenido restringido es aquel que cada uno compartiría con un *Doppelgänger* suyo, entendido como un duplicado exacto del propio organismo, partícula por partícula. La propia Baker define "restringido", entendido como un predicado aplicable tanto a estados como a contenidos, siguiendo la línea inaugurada por Putnam:

> A state is *narrow* if and only if whether or not *x* is that state is determined solely by the properties of *x*, without presupposing that anything other than *x* exists. The belief that *p* has *narrow* content if and only if whether or not *x* believes that *p* is determined solely by properties of *x*, without presupposing that anything other than *x* exists. (Baker 1985: 138)

Ahora bien, la noción de tipo semántico restringido se articula con más precisión sobre la propuesta de Fodor: dos actitudes proposicionales pertenecen al mismo tipo semántico restringido si comparten la proposición siempre que ésta venga expresada en *oratio obliqua* –es decir, de forma indirecta y por ello refractaria a la sustitución *salva veritate* de términos correferenciales–; o bien, equivalentemente, dos actitudes

proposicionales pertenecen al mismo tipo semántico restringido si no hay ninguna diferencia entre las proposiciones que no sea una diferencia de valor de verdad o de referencia a entidades individuales específicas. Pues bien, la cuestión crucial del argumento de Baker es si la individuación de actitudes proposicionales en virtud del tipo semántico restringido al que pertenezcan coincide con su individuación en virtud del tipo de estado funcional con el que se identifiquen. Cualquier respuesta –sostiene Baker– es desalentadora para el funcionalismo: una respuesta negativa nos forzaría a desestimar que las actitudes proposicionales identificadas por su tipo semántico restringido sean estados funcionales, lo que es tanto como rechazar el funcionalismo, pero una respuesta afirmativa nos comprometería con una tríada de afirmaciones incompatibles entre sí, de la que sólo podríamos escapar renunciando, también, a tesis medulares de la concepción funcionalista de lo mental:

> The inconsistent triad is this:
>
> (A) Beliefs individuated by narrow semantic type are psychological states.
> (B) Psychological states are functional states.
> (C) Two tokens of a single functional state may differ in narrow semantic type. (Baker 1985: 139-140)

El primer término de la tríada se desprende directamente de la constatación de que cuando la atribución de actitudes proposicionales interviene en la explicación de la conducta lo hace por lo general de forma opaca, constatación que Fodor ha subrayado a menudo –por ejemplo, en Fodor (1980a: 66). Ahora bien, la individuación de actitudes proposicionales opacas, donde la proposición se expresa en *oratio obliqua*, es por definición un caso de individuación por tipo semántico restringido. Luego las creencias identificadas por tipo semántico restringido son estados psicológicos, en el sentido de que resultan relevantes en la explicación de la conducta. El segundo término de la tríada es, obviamente, la tesis básica del funcionalismo: la afirmación de que, a los efectos relevantes para la explicación psicológica, los estados mentales se taxonomizan por sus relaciones con estímulos, conductas y otros estados mentales[289]. El tercer término de la tríada deriva de lo que, a ojos de Baker, no es más que un hecho frecuentemente reconocido en la teorización cognitiva: que un mismo estado funcional, o secuencia de estados funcionales, es susceptible de múltiples interpretaciones semánticas.

Hacer patente la incoherencia de la tríada requiere algún trabajo lógico. Sean x e y actitudes proposicionales con contenido restringido, $x \neq y$. Entonces, si x es un estado psicológico determinado, P, por (A), y no es ese mismo estado psicológico. Eso a su vez implica que si x es un estado funcional determinado, F, por (B), y no es ese mismo estado funcional. Pero, por (C), podemos suponer en principio la existencia de un x y un y tales que $x \neq y$, y tanto x como y son instancias de un mismo

[289] Que Baker identifica sin más con la idea de Fodor de que la sintaxis constituye el vínculo entre la semántica y la causalidad –*cf. infra*.

tipo de estado funcional, *F*. Tenemos entonces que si *a* es *F*, *b* no lo es, y también que tanto *a* como *b* son *F*, de lo que es sencillo derivar que *b* es y no es *F*[290]. La repercusión de este argumento sobre el funcionalismo cognitivista es, de acuerdo con el diagnóstico de Baker, sumamente severa:

> As long as functional state tokens can have more than one semantic interpretation –and I take it to be a central feature of the computer analogy that they can– mental states like beliefs cannot be understood as functional states. So unless Fodor and other functionalists are willing to abandon the view that beliefs are functional states […], I do not see how they can avoid the contradiction, which stems from the very machine analogy that has given functionalism its impetus. (Baker 1985: 144)

Esquivar la contradicción exigiría afirmar que la individuación de actitudes proposicionales en virtud de su tipo semántico restringido es una forma de individuación insuficientemente restringida para los fines de la explicación cognitiva[291]. Pero esto –insiste Baker– es tanto como renunciar a la idea de que las actitudes proposicionales sean estados funcionales, lo cual a su vez es tanto como renunciar al funcionalismo. Ése es el segundo término del dilema:

> In a word, if the functionalist claims that even beliefs individuated by narrow semantic type are too "wide" to be functional states, there seems to be no beliefs left to be candidates for functional states. (Baker 1985: 145)

Así que, en suma, "[...] either beliefs are not psychological states, or psychological states fail to be functional states. In neither case is the functionalist's optimism borne out" (Baker 1985: 146).

Pero el argumento de Baker queda en entredicho tan pronto como advertimos la debilidad de la tercera de las premisas que forman la tríada de la que se deriva la contradicción. Que distintas instancias del mismo tipo de estados funcionales –de acuerdo, claro está, con los criterios de taxonomización que para dichos estados funcionales auspicia el cognitivismo– puedan diferir en orden a su contenido semántico restringido, lo cual Baker toma –lo hemos visto ya– como un rasgo consabido de la teorización cognitiva o incluso como algo consustancial a la

---

290

(1) $(x \neq y) \rightarrow (Px \rightarrow \neg Py)$	(A)
(2) $(Px \rightarrow \neg Py) \rightarrow (Fx \rightarrow \neg Fy)$	(B)
(3) $(\mathrm{E}x,y)\ (x \neq y\ \&\ Fx\ \&\ Fy)$	(C)
(4) $a \neq b\ \&\ Fa\ \&\ Fb$	Instanciación de (3)
(5) $(a \neq b) \rightarrow (Fa \rightarrow \neg Fb)$	Instanciación de (1) y (2), silogismo hipotético
(6) $(Fa \rightarrow \neg Fb)$	(4), (5), *Modus ponens*
(7) $Fb\ \&\ \neg Fb$	(6), (4), *Modus ponens*, extracción e introducción de la conjunción.

291 Igual que serían insuficientemente restringidas, según Baker (1985: 138, 144), las actitudes deícticas o las actitudes *de re* –*cf.* McDermott (1986, *supra*) para la idea de que las actitudes restringidas atribuidas en la teorización cognitiva son necesariamente actitudes *de re*, pero referidas a estímulos y conductas proximales.

"analogía computacional" (Baker 1985: 144, *supra*), dista en realidad de ser ninguna de esas dos cosas. Antes al contrario, lo que se perfila como una de las tesis medulares de la concepción computacional de la mente es precisamente, como hemos de ver en detalle, que la semántica de los estados internos –si acaso no en lo que incumbe a su referencia o su verdad, sí al menos en su faceta interna, restringida– se ciñe fidelísisamente a su sintaxis –por la que no se entiende otra cosa que su imbricación en patrones causales que involucran estímulos, respuestas y otros estados internos– del mismo modo que sucede cuando comparamos la sintaxis de los procesos computacionales de un autómata con su interpretación semántica. Así pues, un defensor del funcionalismo podría sin reparos, ante el embate de Baker, ceder esa tercera premisa polémica y comprometerse con la tesis de que el contenido semántico restringido de un estado mental guarda una insobornable superveniencia con el patrón de relaciones funcionales que trabe –a cuya luz, por otra parte, lo clasificamos como tal estado mental. Cosa bien distinta es, desde luego, que al hacer tal cosa, el funcionalista asuma una distancia entre nuestra concepción coloquial de lo que es una creencia o un deseo y su propio concepto funcional de representación interna que a la larga acabe por forzarlo a abandonar la idea de que las creencias y los deseos son estados funcionales, aunque no sea, como Baker pretende, porque haya tenido que restringir la noción de contenido sobre la que operan las representaciones internas hasta el punto de que no queden ya en nuestra vida mental trasuntos reconocibles de dichas representaciones, sino, por el contrario, porque la noción de contenido que subyace a la concepción ordinaria de creencias y deseos sea finalmente más abierta de lo que parecía, e incluya inextricablemente nociones como las de referencia o verdad.

Conviene apuntar, por otra parte, que las discrepancias entre la idea de lo semántico que obra en el dominio de la psicología cotidiana y la que haya de articular el despliegue científico del cognitivismo pueden interpretarse de forma que alcancen incluso a la robustez de la propia noción de proposición. Una de las lecturas más influyentes de la controversia entre una concepción internista y una externista del contenido de los estados mentales es, en efecto, aquella según la cual la verdad del externismo desautoriza al concepto de proposición para cumplir el papel que suele atribuírsele en la explicación psicológica –*cf*. Dennett (1982, *supra*). Para desempeñar dicho papel, una proposición debería –recordemos–, además de servir de guía para la acción, ingeniárselas para determinar una extensión y portar un valor de verdad a la vez que resulta aprehensible por la mente. Pero si la extensión determinada por una proposición, y con ella sus condiciones de verdad, pueden variar sin que ello conlleve ningún cambio en el estado interno del sujeto, difícilmente podrá resultar aprehensible por la mente de ese sujeto. Tal como Dennett (1982) condensa la conclusión de Putnam (1975a): "[…] something must give: either meaning 'isn't in the head,' or meaning doesn't determine extension" (Dennett 1982: 11). El propio Dennett remeda el argumento en una depurada versión que merece la pena anotar:

> […S]uppose Twin-Earth is just like Earth except that my wallet is in my coat pocket and my *Doppelgänger*'s wallet isn't in his coat pocket. I believe (truly) that my wallet is in my coat pocket. My *Doppelgänger* has the counterpart belief. His is false, mine is true; his is not about what mine is about –viz., *my* wallet. Different propositions, different propositional attitudes, same psychology. (Dennett 1982: 12)

Ahora bien, dado que en su opinión el razonamiento de Putnam "[…] descansa sobre inciertas doctrinas acerca de los géneros naturales y la designación rígida" (Dennett 1982: 12), es más prudente renunciar a su conclusión, y asumir así alguna suerte de internismo, que aceptarla y vernos forzados a reformular la propia noción de proposición, o las relaciones entre intensión y extensión[292].

Este terco equilibrio en que oscilan los argumentos que nos compelen ora a entender lo mental como inherentemente abierto al mundo, ora a advertir su peculiar reclusión, ha quedado bellamente condensado por Moya (1994: 233-235):

> […L]os estados mentales tienen un determinado contenido. […] Lo peculiar de este contenido es que parece que debe ser a la vez interno y externo, estar a la vez en el sujeto del estado intencional en cuestión y fuera del sujeto, en el mundo. […]
>
> […] Este carácter [interno] de los fenómenos mentales es una de las bases en las que descansa la concepción cartesiana de la mente como un ámbito independiente del mundo externo material, la convicción de que mis creencias, deseos y, en general, mis pensamientos son lo que son con independencia de cómo sea el mundo; es más, seguirían siendo lo que son aun cuando el mundo material no existiera.
>
> Sin embargo, ésta es sólo una parte de la cuestión, porque considerado desde otra perspectiva, lo que creo, deseo o pretendo, el contenido de tales estados, aparece como algo externo a mí. Si, por ejemplo, yo deseo ir al cine, lo que deseo es algo que, de suceder, no tendrá lugar en mi interior, sino en el mundo. Si creo que está lloviendo y compruebo que efectivamente es así, lo que sucede es precisamente lo que yo creía.

La cuestión del error –o, bajo la óptica de los deseos, la constatación, a veces punzante, de que estos no se extinguen por quedar incumplidos– cobra cuerpo entonces, según venimos observando, como la clave que vertebra las exigencias que cabe plantear a nuestra concepción del contenido de creencias y deseos, o acaso de lo mental en general:

> […] Mientras que el carácter interno del contenido destaca con más claridad en los casos de creencias falsas o deseos insatisfechos, su carácter externo se descubre más fácilmente en los casos de creencias verdaderas o deseos satisfechos. Pero en ambos casos estos caracteres se aplican también al resto de creencias y deseos. […]
>
> […U]na concepción correcta del contenido debe poder dar cuenta de ambos aspectos. El internalismo de corte cartesiano tiene dificultades para dar cuenta de la relación de la mente con el mundo externo y en especial del hecho de que el mundo, en ocasiones, hace

[292] Sea como sea, la convicción, alentada por razonamientos externistas, según la cual la noción fregeana de proposición no estaría capacitada para la triple labor que se le exige habría conducido –o así, como hemos visto, piensa Dennett (1982: 8)– al intento de reemplazar a las proposiciones por fórmulas del lenguaje de pensamiento, con la vana esperanza de que éstas estuvieran mejor dotadas para la tarea.

> verdaderas nuestras creencias y satisface nuestros deseos. ¿Cómo algo que es independiente del mundo externo puede hallarse en armonía con él? El externalismo, en cambio, tiene dificultades para explicar la discrepancia entre la mente y el mundo, el hecho de que, en ocasiones, nuestras creencias son falsas y nuestros deseos se frustran. ¿Cómo algo que depende del mundo externo puede estar en discrepancia con él? (Moya 1994: 233-235)

Asumidas, pues, todas estas matizaciones, lo que en el contexto de este trabajo nos preocupa es, una vez más, dejar abierta la pregunta de si una renuncia al individualismo –la ruptura, en otras palabras, con el solipsismo metodológico propugnado por Fodor (1980a)– sería capaz de abrir el cauce del que mane una noción de eficacia causal que podamos verídicamente asignar a lo mental, y que sea suficientemente rica como para legitimar su plena autonomía epistemológica –la irreductibilidad de la explicación psicológica. Es decir: si, dicho de otro modo, cuando se asume que "[…] la mente no es una prisionera encerrada en un recinto umbrío de representaciones, sino que consiste en una serie de funciones que se aplican a los argumentos del medio" (Rivière 2002: 34), esa constatación nos provee además de algún modo de poner en pie la idea de que la mente no es tampoco prisionera del cerebro que la encarna –aun cuando no sea otra cosa que ese cerebro, vivo en ese entorno.

Lingua mentis, recinto umbrío

Acaso como parte de la herencia de Place (1956), donde la discusión sobre la naturaleza del dolor parecía fijar la pauta del debate acerca de lo mental, algunas de las primeras formulaciones del funcionalismo (*cf*. por ejemplo Putnam 1967a, 1967b, 1967c) se desplegaron tal vez desatendiendo la caracterización de las actitudes proposicionales, como creencias o deseos[293]. Con ello, el carácter representacional de de los estados internos del organismo que se proponía identificar, entonces, con los estados de tabla de máquina de una máquina de Turing podía quedar velado. Pero rectificar el rumbo requería, a todas luces, contar con alguna suerte de aparato combinatorio: la productividad y la sistematicidad de las proposiciones que pueden ser objeto de una actitud eran –como ha señalado Block (1996), *cf*. también Block y Fodor (1972a, *supra*)– los principales escollos que era preciso librar.

> In the case of monadic states like pain, the sensation of red, and so on, it does seem a theoretical option to simply list the states and their relations to other states, inputs, and outputs. But for a number of reasons, this is not a sensible theretical option for belief states, desire states, and other propositional-attitude states. For one thing, the list would

[293] Es más: como hemos visto (*cf*. por ejemplo Place 1956: 44, *supra*), el proyecto de Place pasaba por excluir a las actitudes proposicionales de una interpretación fisicalista que se reservaba para las esquivas sensaciones puras; en el ámbito de creencias y deseos, el análisis disposicional de Ryle (1949) aparecía como la aproximación idónea.

> be too long to be represented […]. For another thing, there are systematic relations among beliefs […]. We cannot treat "believes-that-grass-is-green" and "believes-that-grass-is-blue," and so forth as unrelated primitive predicates. (Block 1996: 21)

Lo que necesitábamos para afrontar esas dificultades parecía ser justo, como señala Levin (2004: §4.5), alguna suerte de semántica de rol conceptual, en la que el contenido semántico de una proposición viniera definido por su papel en esa red de relaciones entre ellas: de ese modo, las similitudes y las diferencias en el contenido de nuestras creencias y deseos podrían entenderse como similitudes y diferencias en las proposiciones hacia las que mantenemos la actitud de creer o desear, mientras que mantener una determinada actitud hacia tal o cual proposición se entendería a su vez como albergar un estado cuyas relaciones causales con otros estados internos que también constituyen actitudes proposicionales reflejan las relaciones semánticas entre las proposiciones pertinentes. Trabajos como los de Field (1980), Loar (1981) o Block (1986) cobraban entonces un papel decisivo en la fundamentación del funcionalismo, proporcionádole una semántica afín.

También se pregunta Levin (2004: §4.5), casi al vuelo, en qué medida una semántica de rol conceptual para las actitudes proposicionales puede pasar sin postular algo así como un lenguaje interno de la mente. Las mismas premisas que señala Block (1996: 21) sustentan, en efecto, buena parte de los argumentos que llevaron a Fodor (1975) y Field (1978) a elaborar la hipótesis de que las operaciones computacionales descritas por una teoría psicológica madura habrían de efectuarse sobre un código simbólico capaz de expresar de forma estructurada las proposiciones que dan contenido a nuestras actitudes, un código que vendría a constituir –por tanto– el lenguaje del pensamiento[294]. Que el número de las creencias o los deseos que podemos abrigar es ilimitado, y que el conjunto de las que de hecho alberguemos tiende a exhibir una notable estructura interna, se convertían así en los motivos cruciales para entender los procesos psicológicos bajo la imagen del lenguaje interior. No en vano, la composicionalidad que se precisaba otorgar a dichos procesos parece encontrarse –como nos recuerdan Fodor y LePore (1991: 146):

> […] at the heart of some of the most striking properties that natural languages exhibit. The most obvious of these are *productivity* (roughly, the fact that every natural language can express an open-ended set of propositions) and *systematicity* (roughly, the fact that every natural language that can express the proposition P will also be able to express many propositions that are semantically close to P. If, for example, a language can express the proposition that aRb, then it can express the proposition that bRa […].)

Además, la idea de un lenguaje del pensamiento estaba llamada a proveer a la explicación psicológica de una distinción entre semántica y sintaxis que había de

[294] Puede verse una somera introducción a la hipótesis del lenguaje del pensamiento, entre tantas, en Hermoso (2001). Abordarla en detalle, por no hablar de aquilatar su fundamentación o sus consecuencias en diversos órdenes, exigiría una investigación cuando menos tan detenida como la que nos ocupa, cuyos límites, desde luego, desborda con mucho.

resultar crucial para su articulación en una teoría computacional. A tal efecto, lo que se entresacaba de los intricados vínculos entre propiedades semánticas y sintácticas era –por motivos en los que hemos de profundizar más adelante– una idealizada cohesión:

> Connected to both productivity and systematicity is a further, apparently perfectly universal, feature of natural languages. The structure of sentences is, in the following sense, *isomorphic* to the structure of the propositions they express: *if a sentence S expresses the proposition that P, then syntactic constituents of S express the constituents of P*. If, for example, a sentence expresses the proposition that P and Q, then there will be one syntactic constituent that expresses the proposition that P and another syntactic constituent that expresses the proposition that Q. (Fodor y LePore 1991: 147)

Ante la constatación, así pues, de que las creencias o deseos que llegamos a albergar a lo largo de la vida no son sino una ínfima parte de la enorme variedad que habrían podido dársenos, y dado que un mínima lealtad al naturalismo nos impide dejar sin explicar la relación entre las infinitas proposiciones que podríamos, por ejemplo, creer y las propias creencias[295], la conclusion natural –piensa Fodor– es que:

> […] the productivity of thoughts is like the productivity of natural languages, i.e., there are indefinitely many thoughts to entertain for much the same reason that there are indefinitely many sentences to utter. (Fodor 1985: 22)

Pero entonces:

> The moral for treatments of the attitudes would seem to be straightforward: solve the *productivity* problem for the attitudes by appealing to constituency. Solve the *constituency* problem for the attitudes in the same way that you solve it for speech-acts: tokening an attitude involves tokening a symbol, just as tokening an assertion does. What kind of symbol do you have to token to token an attitude? A mental representation, of course. Hence [the] R[epresentational] T[heory of] M[ind]. (Fodor 1985: 23)

Una mirada tan penetrante como recelosa sobre las raíces de la idea de la *lingua mentis* –y acaso, por eso mismo, particularmente reveladora– puede encontrarse en Dennett (1982, 1991a). El vigor de los argumentos que nos conducen a la hipótesis del lenguaje del pensamiento reside según Dennett, efectivamente, en la dificultad de hallar explicación para la capacidad de manejar descomunales cantidades de datos exhibida por el sistema nervioso humano, si no es apelando a "principios de

[295] Dicho de otro modo:

> Since relations between organisms and propositions aren't to be taken as primitive, one is going to have to say what it is about organic states like believing and desiring that allows them to be (roughly) as differentiated as the propositions are. If, for example, you think that attitudes are mapped to propositions in virtue of their causal roles […], then you have to say what it is about the attitudes that accounts for the productivity *of the set of causal roles*. (Fodor 1985: 22)

representación elegantes, *generativos* e indefinidamente extensibles", principios para los que el único modelo conocido es el lenguaje: –"[…s]o the argument for a language of thought comes down to this: what else could it be? (Dennett 1991a: 329). Ya en *El lenguaje del pensamiento*, desde luego, el propio Fodor había dejado claro que la primera premisa de su argumentación era que "[…] the only psychological models of cognitive processes that seem even remotely plausible represent such processes as computational", y nos comprometen, por tanto, a atribuir a los organismos o los autómatas a los que hayan de aplicarse dichos modelos "[…] a medium of computation: a representational system" que sólo podría ser, a su juicio, un lenguaje interno (Fodor 1975: 27). Más contundente es la expresión que Lycan (1993: 407) daría a esta convicción de que la concepción computacional de la mente es, como suele decirse, *the only game in town*[296]: "[…] could the thing be otherwise, barring either magic or divine intervention?"

El propio Dennett (1982) había desplegado un análisis mucho más minucioso de las razones por las que la hipótesis del lenguaje del pensamiento ejerce sobre el pensamiento la seducción –perversa, a su entender– que al parecer ejerce. La cuestión sería –en esencia– que convertir la idea de que alguien mantenga una actitud hacia una *proposición* en la idea de que alguien mantenga una actitud hacia una *oración* nos permite esquivar algunas de las notorias dificultades que afligen a la noción de actitud proposicional[297]. Esas notorias dificultades radican en la constatación, que debemos a Frege (1892), de que las proposiciones de nuestra teoría de las actitudes proposicionales –sus "Pensamientos"– deben simultanear tres condiciones: la de ser portadoras de un valor de verdad, la de ser determinantes de una extensión (o sea: la de constituir una intensión), y la de ser aprehensibles por la mente[298]. A ello se suma, según Dennett, la confusa idea de que las proposiciones deban, para poder participar en la explicación psicológica, desempeñar algún papel causal en nuestra vida psíquica[299]. Así visto, resulta claro que:

---

[296] *Cf.* en Fodor (1985: 27, *infra*), otro de los *loci classici* de tal esquema argumental, una severa crítica de cuya legitimidad puede hallarse en Achinstein (1990).

[297] Para resguardarse del canto de la sirena, Dennett –como Ulises que tapara con cera sus oídos–se adhiere a la interpretación deflacionista de Churchland (1979), de acuerdo con el cual "*x* cree que *p*" es el mismo tipo de expresión que "*x* se mueve a 5 m/s", la cual a su vez es el mismo tipo de expresión que "x se mueve rápidamente". Si bien el análisis adverbial de Churchland –admite Dennett– no resuelve definitivamente las perplejidades metafísicas desatadas por la noción de proposición, al menos nos mantiene provisionalmente firmes ante la invocación del lenguaje del pensamiento.

[298] Es inevitable que las condiciones que dejó fijadas Frege evoquen a las que Block (1986: 87, *supra*) trataría de imponer a aquella semántica que hubiera de servir de sustento a la psicología: que diera cuenta de la relación entre la noción restringida y la noción amplia de significado, que aclarase cómo y por qué cada una de ellas es relevante en la explicación psicológica, y que nos dejara ver cómo puede el significado restringido determinar una referencia, o un valor de verdad, dado un contexto.

[299] A este respecto, *cf. infra* sobre los aprietos que las observaciones antiindividualistas de Bruge (1986) suponen, de acuerdo con Dennett (1982) para una psicología que quiera construirse sobre la noción de actitud proposicional: si la extensión determinada por una proposición, y con ella sus condiciones de verdad, pueden variar sin que ello conlleve ningún cambio en el estado interno del sujeto –como sostiene Burge–, difícilmente podrá resultar aprehensible por la mente de ese sujeto. Según esto, el

> For propositions to have such a function, they must be concrete –or have concrete tokens– and this leads inevitably to a version of [the] view […] [that] propositions are sentence-like entities. (Dennett 1982: 8)

El intento de dar algún sentido a la noción de aprehensión de una proposición de tal forma que su aptitud para guiar la conducta quede intacta es, entonces, una de las rutas que desembocan en la idea de una *lingua mentis*. Hay otras tres que Dennett intenta cartografiar: padecemos, en primer lugar –lo hemos visto– cierto estupor ante la composicionalidad exhibida por las actitudes proposicionales, que nos llevaría a buscar sus *elementos* en símbolos discretos, lingüísticos o para-lingüísticos; con tal estupor cohabitaría el anhelo de una explicación de la opacidad referencial como la que parece proporcionarnos la idea de que Edipo pueda desear casarse con Yocasta, pero no con su propia madre, *porque* "Yocasta" y "la madre de Edipo" son símbolos sintácticamente diferenciados –*ergo* físicamente diferenciados– en la cifra de su mente; por último, las fórmulas de un código interno –como las oraciones de cualquier lengua– podrían incorporar elementos deícticos a los que las proposiciones son "impermeables" (Dennett 1982: 15), concediéndonos así ni más ni menos que "[…] el trasunto lingüístico de la relatividad a un punto de vista subjetivo que constituye el marchamo de lo mental" (Dennett 1982: 16). Por una senda u otra, al fin y al cabo: "[…] the "expressions" in the language of thought are needed as the "raw material" for psychologico-semantic interpretation of psychological states" (Dennett 1982: 20).

Cuando ya había repudiado el cognitivismo cuyos fundamentos tan decisivamente había contribuido a asentar, Putnam (1985: 211) atribuiría a la concepción de lo mental defendida por Fodor tres dogmas independientes entre sí: uno, que los procesos computacionales que la mente efectúa se desarrollan en un lenguaje con estructura cuantificacional fregeana; dos, que la explicación psicológica cotidiana en términos de creencias y deseos describe de hecho operaciones en dicho formalismo, y tres, que el contenido de tales creencias y deseos corresponde a las mismas expresiones del formalismo con independencia del lenguaje natural del sujeto. En el seno de lo que Dennett (1986: *passim*) venía denominando la "Alta Iglesia Computacional", cuyo sumo pontífice sería Fodor, la trinidad dogmática quedaba descrita en otros términos –aunque parcialmente coincidentes. En cualquiera de las dos reconstrucciones de la ortodoxia cognitivista, la hipótesis –o el dogma– del lenguaje del pensamiento aparece como el eje en torno al cual rota todo lo demás:

---

externismo inhabilitaría al concepto de proposición para la tarea que Frege le encomienda, lo cual de nuevo sería a ojos de Dennett un motivo, aunque espurio, para reemplazar las proposiciones por fórmulas del lenguaje del pensamiento.

> (1) *Thinking is information processing*. […T]he terms of folk psychology are to be spruced up by the theorist and recast more rigorously […].
> (2) *Information processing is computation (which is symbol manipulation)*. […A] medium of representation is posited, consisting of *symbols* belonging to a *system* which has a *syntax* (formation rules) and *formal rules of symbol manipulation* for deriving new symbolic complexes from old.
> (3) The semantics of these symbols connect thinking to the external world. […T]he symbolic structures composable within the representational medium have interpretations that are a systematic function of the semantic interpretations of their elements. In other words, there is a language of thought […]. (Dennett 1986: 60-61)

Estos principios coincidían en lo esencial, como concisamente se encarga de recordarnos Haugeland (2002a), con los del proyecto de investigación que John McCarthy –en McCarthy, Minsky, Rochester y Shannon (1955)– había bautizado con el nombre de "inteligencia artificial". A saber:

> […] "thinking" ("reasoning", "problem solving") takes the form of rule-governed manipulations of broadly sentence-like internal tokens. Their sentence-likeness consists in this: (1) they are composites of atomic units (drawn from a finite prespecified list of atoms) and formed according to finitely many prespecified recursive formation rules; (2) each well-formed token has a semantic interpretation that is fully (and recursively) determined by its form and the atoms of which it is composed; and (3) the manipulation rules are designed so as to preserve, on the whole, various semantic desiderata (such as truth, probability, and/or conduciveness to goal achievement). (Haugeland 2002a: 24-25)

Como tantos otros, también Block (1980a: 28) había señalado que este compromiso con la idea de una *lingua mentis* vertebra la concepción de lo psicológico del funcionalismo cognitivista:

> Psychological states are seen as systematically representing the world via a language of thought, and psychological processes are seen as computations involving these representations. (Block 1980a: 28)

La convicción de que si las creencias o las opiniones son acerca de los hechos que las harían verdaderas, como los deseos, los temores o los anhelos son acerca de los hechos que los harían cumplirse, es *porque* unos y otros *representan* la realidad de cierta manera, y lo hacen en el lecho de ese lenguaje interior, aparece, así pues, como un rasgo consustancial al cognitivismo. Como apunta concisamente García-Carpintero (1995: 44):

> Si los estados mentales tienen contenido es que ellos también *representan* la realidad de un cierto modo: el contenido de los estados mentales consiste en la especificación de los detalles de la representación que en ellos se hace de la realidad.

En un gesto de marcado aire brentaniano –recordemos que, a ojos de Brentano, "[…n]ada puede ser juzgado, nada tampoco apetecido, nada esperado o temido, si no es representado" (Brentano 1874: 13), o, más en general, que "[…g]racias a la

generalidad con que usamos la palabra, pudimos decir que es imposible que la actividad psíquica se refiera a algo que no sea representado" (Brentano 1874: 90)–, queda así identificada la relación de intencionalidad que enlaza a un estado psicológico con su contenido, o a una actitud proposicional con su proposición, con la relación de representación. Pero que esas representaciones que conforman nuestros juicios, apetencias, esperanzas o temores no sólo sean todas de la misma naturaleza –ideas– sino que puedan engarzarse entre sí, que de hecho suelan estar formadas por combinaciones de ideas elementales que operan a modo de átomos del pensamiento, es desde luego –como se ha señalado con frecuencia[300]– una tesis que no arraiga en Brentano sino en Locke.

Sería hermoso trazar el recuento de cómo, desde Locke, la idea de que el pensamiento –o acaso toda la vida mental– pueda desplegarse en un lenguaje propio proviene en último término del descubrimiento en uno mismo de un soliloquio silencioso en el mismo lenguaje en que uno habla en alta voz: así en las deliberaciones que atormentan a Agenor junto a los muros de Troya cuando repara en que debe enfrentarse al temible Aquiles y, mientras "[…] su agitado corazón vacilaba sobre el partido que debería tomar", habla, "[…] gimiendo, a su magnánimo espíritu" (*Ilíada* XXI. 544), o también en el "[…] diálogo interior y silencioso del alma consigo misma" del que el extranjero habla a Teeteto en *Sofista* 263e, asemejándolo al "[…] flujo que surge de ella y sale por la boca, acompañado de sonido"-*cf*. también *Teeteto* 189*e*, 206*d*–, o en el *sermo interior* en que Agustín de Hipona cree intuir el trasunto en nuestra alma del Verbo de Dios: "[…] la palabra que brilla dentro, a la cual responde mejor el nombre de 'palabra'" (*De Trinitate* XV, 11, 20), y que se hace carne en nuestra voz, "[…] pues aunque las palabras no suenen, quien las piensa las dice ciertamente en su corazón" (*De Trinitate* XV, 10, 17).

Desde luego, entre los cauces por los que habría de discurrir ese relato se contaría la descripción del modo en que la *lingua mentis* va apartándose del lenguaje natural con el que inicialmente se identifica: cómo, por ejemplo, va articulándose la constatación de que multitud de organismos, aun careciendo de un lenguaje natural, exhiben capacidades perceptivas complejas, así como una insoslayable capacidad de aprender y de orquestar medios con vistas a fines, y a partir de esa constatación, cómo se construye la idea de que un lenguaje interno debe sustentar los procesos de resolución de problemas mediante formulación y comprobación de hipótesis que se postulan para explicar no sólo dichas capacidades sino también, cuando se da, la adquisición de la propia lengua natural[301]. Pero por otra vertiente de ese mismo

[300] *Cf*. por ejemplo Cummins (1989: 157), que atribuye a Locke un compromiso velado con la noción de lenguaje del pensamiento.

[301] Estos son, de hecho, los principales argumentos que Fodor (1975: 30-44) esgrime para justificar la postulación de un código interno de representación distinto del lenguaje natural; Patricia Churchland (1978), en cambio, interpreta los mismos datos que Fodor aduce como indicio primordial de la existencia de la *lingua mentis* –las capacidades cognitivas de organismos no-lingüísticos– como la base de una *reductio ad absurdum* de la posición de Fodor.

relato transcurriría, sin duda, el esfuerzo por entender las razones en virtud de las cuales, a partir seguramente de los estrechos lazos que Descartes trazara entre lo mental y lo que se presenta a la consciencia, ese diálogo del alma consigo mismo deja de ocurrir, aunque *aneu phonés*, a plena luz del día, y comienza a verse como algo que tiene lugar en un recinto cerrado al que, como en una *camera obscura*, el mundo llega sólo a través del angosto conducto que abren los sentidos. No en vano advertía ya John Locke (1690: II, XI; §17) en su *Ensayo* que:

> […] las sensaciones exteriores e interiores son las únicas vías por donde yo encuentro que el conocimiento llega al entendimiento. Hasta donde alcanzo a descubrir, éstas son las únicas ventanas por donde pueda entrar la luz a ese *cuarto obscuro.* Porque, paréceme que el entendimiento no es muy desemejante a un gabinete completamente oscuro, que no tendría sino una pequeña abertura para dejar que penetraran las semejanzas externas visibles, o si se quiere, las ideas de las cosas que están afuera; de tal manera que, si las imágenes que penetran en un tal cuarto oscuro pudieran quedarse en él, y se acumularan en un orden como para poder ser encontradas cuando lo pida la ocasión, habría un gran parecido entre ese cuarto y el entendimiento humano […].

Si la evocación de aquel vetusto cuarto obscuro había de resurgir en la caja negra de los modelos conductistas, no está de más anotar que el auge de la psicología cognitiva y el funcionalismo, bajo el aire solipsista de su énfasis en el papel desempeñado por la representación interna en la determinación de la conducta y la insistencia en ligar el contenido semántico de tales representaciones, además de a sus relaciones mutuas, a las que trabaran con estímulos y conductas entendidos en sentido proximal, parecería haber alumbrado una *caja blanca*: un espacio resplandeciente enclaustrado en un mundo que le es ignoto. La *camera obscura* de la mente, por oscura que sea, sería entonces el único ámbito luminoso: fuera de ella no habría ya sólo oscuridad, sino ceguera[302]; sería, acaso como la consciencia cartesiana, una habitación de paredes opacas pero interior transparente.

La metáfora de la llave y la soberanía del significado

Tan pronto como comenzamos a pensar acerca de nuestras propias creencias y deseos bajo el prisma del lenguaje –incluso si no supusiéramos que éstas se expresen en ese código interno que los psicólogos habrían de ir desentrañando, sino en nuestra propia lengua materna–, comienza a tomar forma una pregunta que acabaremos

---

Que el lenguaje del pensamiento debe, si es que ha de servir para las tareas explicativas que se le encomienda, depurar las innumerables ambigüedades que erizan cualquier lenguaje natural (*cf*. por ejemplo, Block 1986: 120, *infra*) es otro motivo para desligarlo de estos.

[302] Dicho de otro modo: ver la mente como una caja negra exigía en realidad extraer de ella cuanto en su transparencia se nos muestra e imponérselo al mundo, forzando a cada paso las nociones de estímulo y respuesta para cargarlas del repudiado vocabulario mentalista: esa habría sido, bajo esta óptica, la sustancia última de las críticas de Chomsky (1959) contra Skinner (1957).

teniendo que abordar: ¿en qué medida están involucradas en la determinación de nuestra conducta, o de la sucesión de nuestros pensamientos, afectos o deseos, las propiedades semánticas que esa perspectiva lingüística hace visibles? Cabría pensar que aquello a lo que se refiere una creencia o un deseo, el hecho que creemos o deseamos –o tal vez mejor: el hecho de que la creencia o el deseo se refieran a ello–, así como la manera en que esa referencia quede anclada, el sentido en que se tracen los vínculos que con los que unzamos los hechos creídos o deseados, sin duda han de ser el norte de toda ramificación que de la creencia o el deseo en cuestión pueda germinar ya en nuestra conducta, ya en nuestra propia vida mental. Lo que rige, entonces, los efectos que un estado mental pueda desplegar no es, fundamentalmente, otra cosa que su significado. Cabe pensar también, no obstante, que para averiguar qué eco pueda provocar en la mente o la conducta de un sujeto una cualquiera de sus creencias o sus deseos –entendidos, ya se sabe, como expresiones lingüísticas con la que el sujeto mantiene una determinada relación– bastaría con atender a las propiedades sintácticas de dichas expresiones; de su significado, así vistas las cosas, nada nos impediría hacer caso omiso.

Ésta parece ser la encrucijada en la que, como Fodor antes que él, se viera ya Field (1978), que supo ligar estrechamente la cuestión al propósito que alentara nuestras investigaciones:

> I have said that the *syntax* of a system of internal representation should be explicitly stated in a psychological theory of belief and desire. Should the *semantics* of the system of internal representation *also* be stated as part of the psychological theory? That depends on what we want psychological theory for. If the task of psychology is to state
>
> (i) the laws by which an organism's beliefs and desires evolve as he is subjected to sensory stimulation, and
> (ii) the laws by which those beliefs and desires affect his bodily movements,
>
> then I think it is clear that we do not need to use the semantics of the system of internal representation in stating the psychological laws: the sentences in the system of internal representation might as well be meaningless as far as the psychology is concerned. (Field 1978: 62)

Que pudiéramos permitirnos ignorar la semántica –siempre que asumiéramos, claro, esta delimitación de los fines de la psicología, que incluye la reformulación de la idea de conducta que Burge (1986, *supra*) denunciaría– era una consecuencia natural del modo en que, de acuerdo con el funcionalismo de inspiración computacional, se articulaba la sintaxis del sistema de representaciones internas que se tomaba como núcleo de la vida mental:

> […W]e have seen that the syntax and type-identity conditions for a system of internal representation should be regarded as functionally characterized by a psychological theory in which they appear; and we can take that theory to be narrow psychology, that is, the kind of psychology that does not employ any semantic characterizations of the sentences in a system of representation. This is important, for it means that the syntax

> and conditions of type-identity for the system of representation could in principle be determined independently of any considerations about what the sentences in the system mean. (Field 1978: 62)

Los primeros intentos de construir modelos mecánicos que remedaran ciertos aspectos de la conducta animal o humana –lo hemos visto de la mano de Cordeschi (2002)– anteceden a la invención de la computadora digital y, *a fortiori*, de los lenguajes de programación. No era infrecuente que al reflexionar sobre aquellos autómatas aflorase la cuestión de que determinados estados físicos de su maquinaria constituyeran una representación de ciertas propiedades del entorno en el que el artefacto se desenvolvía –la idea, como se verá, está ya presente en Lotka (1925). Sin embargo, la pregunta de si es esa propiedad de índole representacional lo que determina el efecto del estado físico en cuestión sobre el comportamiento del autómata ni siquiera se plantea: habría sido sencillamente descabellado pensar que ninguna cosa distinta de las características hidráulicas, eléctricas, magnéticas o, en fin, mecánicas de las piezas cuidosamente ensambladas por el ingeniero –como no fueran, claro, factores ambientales– pudiera pesar sobre el funcionamiento de su flamante criatura. Es más: comoquiera que el carácter representacional que pudiera atribuirse a ciertos estados de la máquina no venía acompañado de otras propiedades inequívocamente lingüísticas, la noción de que tales estados físicos poseyeran una sintaxis y una semántica tardaría en abrirse paso.

De hecho, en la mayoría de las "imitaciones de la vida" (*cf*. Grey Walter 1950, *supra*) que precedieron al auge de la simulación computacional de procesos psicológicos eran precisamente las propiedades sintácticas lo que faltaba, pues si bien la posición de –supongamos– una determinada rueda dentada podía tomarse como representación de un hecho determinado, eran pocas o nulas las posibilidades de combinar dicha posición con otras propiedades físicas de la máquina para construir representaciones de otros hechos, distintos o, menos aún, más complejos. La capacidad combinatoria era, en cambio, una de las virtudes esenciales de los autómatas matemáticos, como las máquinas de Babbage, pero en esos casos, puesto que no había representación de hechos ambientales que guiara la conducta de la máquina, sino representación de expresiones matemáticas –de otras representaciones, pues– que el operador trabajosamente había de codificar, eran las propiedades semánticas de las ruedas dentadas lo que quedaba velado; además, al hallarse los números o los signos inscritos sobre las propias piezas, la apariencia de la máquina era propicia a que las propiedades semánticas de sus estados físicos se dieran por sentadas y, en consecuencia, pasaran desapercibidas. Por un motivo o por otro, en suma, la idea de que los estados físicos de determinadas máquinas conformaran una suerte de lenguaje, con su sintaxis y su semántica, quedaba desdibujada en, siquiera, un vago recuerdo de la idea de *characteristica universalis* y su raigambre lulliana.

Una vez los viejos y acaso ya chirriantes autómatas mecánicos fueron siendo desplazados de los laboratorios por pulcros programas informáticos, la noción de

código comenzaría a ocupar el primer plano y la distinción entre propiedades sintácticas y semánticas de dicho código se impondría por si sola. Desde sus primeros pasos con el *Logic Theorist* y el *General Problem Solver*, Allen Newell, Herbert Simon y J. Cliff Shaw se acostumbraron a tratar los símbolos de la lógica formal, e incluso los enunciados que estos conformaban, como el rigor formalista exige: sin tomar en consideración su significado –es decir, atendiendo únicamente a su aspecto formal, o, como luego recordarían Simon y Siklóssy (1972: 2), "[…] as though [they] were […] pieces of wood or metal". Ya en Newell, Shaw y Simon (1957: 129), por ejemplo, se nos advertía de que:

> Normally the variables of the system [of sentential calculus] are interpreted as sentences, and the axioms and rules of inference as formalizations of logical operations, *e.g.*, deduction. However, L[ogic] T[heorist] deals with the system as a purely formal mathematics, and we will have no further need of the interpretation.

Apenas un año después, en noviembre de 1958, los asistentes al Simposio sobre Mecanización de los Procesos de Pensamiento que tuvo lugar en el Laboratorio Nacional de Física de Teddington, al suroeste de Londres, verían como Marvin Minsky articulaba su defensa del uso de computadores digitales para construir modelos del funcionamiento cerebral, que Frank Rosenblatt rechazaba, en torno a la idea de que la programación parte del estudio de "[…] syntactic processes involving the manipulation of symbolic expressions" (Minksy 1959: 24 *apud* Cordeschi 2002: 189).

En la noción de sintaxis que el cognitivismo comenzaba a forjar confluyeron, así, una idea de gramática formal heredada de la lógica a través de los incipientes ensayos de simulación computacional del razonamiento y una idea de causalidad que provenía de la familiaridad con los primeros artefactos mecánicos diseñados para remedar algún aspecto de la conducta de los organismos. Tal noción de sintaxis, de este modo, quedó ligada a la idea de reglas de manipulación de símbolos enteramente ciegas al significado de dichos símbolos, por un lado, y a la de una concatenación enteramente local de causas y efectos, como en el giro de una rueda dentada. Que el trasiego material o figurado de ruedas y palancas preservaría las propiedades semánticas relevantes se asumió como algo que debería ocurrir sin que esas propiedades semánticas interviniesen en el proceso, tal como ocurría ya en las viejas máquinas de cálculo de Babbage o, a duras penas, en la de Pascal[303].

La idea de que la semántica se plegará a la sintaxis como un reflejo en aguas mansas es, en definitiva, uno de los contrafuertes del cognitivismo. La eficacia causal se entiende como privativa de la sintaxis, y ésta a su vez como una propiedad física relacionada con la forma –en sentido enteramente literal– de los símbolos; la

[303] Tal vez resulte interesante advertir que el sentido que se confiere a lo formal se invierte así del todo respecto al que tuvo en la distinción escolástica entre *supossitio formalis* –el uso de la palabra para aludir a su referente– y la *suppositio materialis* –la mención de la palabra como acto reflexivo–, *cf. supra*.

semántica parece perfilarse como nada más que un eco inerte de la sintaxis. Así, en Fodor (1985) –acaso el *locus classicus* de esta idea:

> Computers show us how to connect semantical with causal properties *for symbols*. So, if the tokening of an attitude involves the tokening of a symbol, then we can get some leverage on connecting semantical with causal properties *for thoughts*. Here, in roughest outline, is how the story is supposed to go.
>
> You connect the causal properties of a symbol with its semantic properties via its syntax. The syntax of a symbol is one of its second-order physical properties. To a first approximation, we can think of its syntactic structure as an abstract feature of its (geometric or acoustic) *shape*. Because, to all intents and purposes, syntax reduces to shape, and because the shape of a symbol is a potential determinant of its causal role, it is fairly easy to see how there could be environments in which the causal role of a symbol correlates with its syntax. It's easy, that is to say, to imagine symbol tokens interacting causally *in virtue of* their syntactic structures. The syntax of a symbol might determine the causes and effects of its tokening in much the way that the geometry of a key determines which locks it will open. (Fodor 1985: 26)

Apenas unos renglones después, de hecho, Fodor alude explícitamente a la influencia de la teoría de la prueba –el esfuerzo, inspirado en el trabajo de Hilbert, por presentar toda demostración lógica o matemática, o prueba, como un objeto formal susceptible de ser sometido a análisis y verificación mediante técnicas exclusivamente formales– en esta concepción de la relación entre semántica, sintaxis y causalidad bajo la metáfora de la llave y la cerradura[304]:

> […W]e know from formal logic that certain of the semantic relations among symbols can be, as it were, "mimicked" by their syntactic relations; that, when seen from a very great distance, is what proof-theory is about. So, within certain famous limits, the semantic relation that holds between two symbols when the proposition expressed by the one is implied by the proposition expressed by the other can be mimicked by syntactic relations in virtue of which one of the symbols is derivable from the other. We can therefore build machines which have, again within certain limits, the following property: the operations of such a machine consist entirely of transformations of symbols; in the course of performing these operations, the machine is sensitive solely to syntactic properties of the symbols; and the operations that the machine performs on the symbols are entirely confined to alterations of their shapes. Yet the machine is so devised that it will transform one symbol into another if and only if the symbols so transformed stand in certain *semantic* relations; e.g., the relation that the premises bear to the conclusion in a valid argument. Such machines –computers, of course– just *are* environments in which the causal role of a symbol token is made to parallel the inferential role of the proposition that it expresses.
>
> […] Computers are a solution to the problem of mediating between the causal properties of symbols and their semantic properties. (Fodor 1985: 26-27)

[304] Que Dennett, en un ejercicio de funambulismo en el que parece deleitarse con especial frecuencia en *Kinds of Minds –cf.* Hermoso (2001b)–, jugaría a interpretar literalmente: "A lock and key exhibit the crudest form of intentionality" (Dennett 1996: 35).

Ya años atrás, cuando emprendiera su influyente defensa de la tesis de que la explicación psicológica debe operar bajo un estricto solipsismo metodológico –es decir, obviando por principio cualquier variable externa a la psicología del sujeto, entendiendo por tal toda aquella que no esté representada en su mente–, Fodor había fijado la posición que, respecto a la eficacia causal del contenido de los estados mentales, parece haberse erigido como ortodoxia en el seno del cognitivismo, y que se describía allí como una consecuencia lógica del hecho de que los procesos mentales sean procesos computacionales: "[…] mental processes have access only to formal (nonsemantic) properties of the mental representations over which they are defined […]" (Fodor 1980a: 63). Dicho de otro modo, "[…] the computational theory of mind requires that two thoughts can be distinct in content only if they can be identified with relations to formally distinct representations" (Fodor 1980a: 64)[305]. Se trata, pues, de una condición de superveniencia de (la identidad entre) pensamientos sobre (la identidad entre) representaciones formalmente descritas con las que dichos pensamientos guardan una determinada relación.

Las ligaduras que atan entre sí, en el núcleo del cognitivismo, a la concepción formalista de los procesos psicológicos y al solipsismo metodológico –ligaduras que se urden, como veremos, al abrigo de la intensionalidad del discurso psicológico ordinario acerca de creencias y deseos– quedan al descubierto en la formulación que ofrece McGinn (1982: 208) de esta *condición de formalidad* fodoriana:

> […] beliefs play a role in the agent's psychology just in virtue of intrinsic properties of the implicated internal representations –the semantic *relations* between representations and things in the world must be irrelevant to the psychological role of beliefs. More precisely, the causal role of a belief must depend upon, and only upon, those properties of representations that can be characterized without adverting to matters lying outside the agent's head.

Así que, si como argumentara Putnam (1975a), el significado *no está en la cabeza* –en la medida en que depende crucialmente de hechos mundanos–, entonces el significado no es, de acuerdo con el cognitivismo, psicológicamente relevante (*cf*. Searle 1992: 203, *supra*). No menos claras son, en este sentido, las palabras de LePore y Loewer (1989: 165), que de hecho citan expresamente a McGinn (1982) justo después de apuntar que:

[305] La condición de formalidad no parece difícil de hacer encajar con la primera de las tres tesis básicas del funcionalismo que bosquejara Baker (1985: 139-140, *supra*) –a saber, que las creencias, individualizadas según su tipo semántico restringido, son estados psicológicos–, ni tampoco con la segunda –que los estados psicológicos son estados funcionales–, pero no está claro que pueda conciliarse con la tercera –que dos casos del mismo tipo de estado funcional pueden pertenecer a distintos tipos semánticos restringidos. No en vano, como se apuntaba entonces, es precisamente la tercera de esas tesis a la que el funcionalista sería más proclive a renunciar para hacer frente a los argumentos de Baker.

> The mind (and its components) has no way of recognizing the reference or truth conditions of the representations it operates on. Instead, it operates on syntactic features of representations which "represent" the semantic features.

No en vano, es en los primeros compases de una de sus vigorosas críticas de los presupuestos individualistas de la psicología cognitiva, donde Burge (1986, *supra*) da fe también, somera pero diligentemente, de la ubicuidad de lo intencional en el discurso teórico de la disciplina:

> [Psychology] […] accepts, for example, that people see physical objects with certain shapes, textures, and hues, and in certain spatial relations, under certain specified conditions. And it attempts to explain in more depth what people do when they see such things, and how their doing it is done. Psychology accepts that people remember events and truths, that they categorize objects, that they draw inferences, that they act on beliefs and preferences. And it attempts to find deep regularities in these activities, to specify mechanisms that underlie them, and to provide systematic accounts of how these activities relate to one another. In describing and, at least partly, in explaining these activities and abilities, psychology makes use of interpreted that-clauses and other intensional constructions –or what we might loosely call "intentional content". (Burge 1986: 43)

Una dilucidación más prolija de la condición de formalidad que operaría según Fodor sobre los procesos psicológicos, así como de su vinculación con la noción de superveniencia y con las tesis solipsistas, nos la proporcionan LePore y Loewer:

> Fodor claims that a consequence of the C[omputational] T[heory of] M[ind] is a formality condition, which specifies that in the C[omputational] T[heory of] M[ind] psychological states count as different states only if they differ computationally. […] This supervenience principle, that *S* and *S** are distinct psychological states only if they are distinct computationally, lies at the heart of the C[omputational] T[heory of] M[ind]. Although Fodor endorses the formality condition, he also thinks that cognitive psychology contains true generalizations connecting propositional attitudes with each other, environmental conditions, and behavior. […] As Fodor emphasizes, the specification of propositional content in these generalizations is essential to their explanatory role. […] At first, this may seem incompatible with the claim that only formal properties of representations are relevant to the computations which produce behavior. However, there is no incompatibility as long as the contents of attitudes are specified in a way that respects the formality condition. […] Fodor observes that a characterization of meaning which conforms to the formality condition is methodologically solipsistic (Putnam […] 1975[a]) in that differences of meaning depend entirely upon internal mental characteristics, *e.g.*, computations over representations. (LePore and Loewer 1989: 165-166)

La misma idea cristaliza en la depurada sinopsis de las tesis cognitivistas –más exactamente: de la concepción de los procesos psicológicos como computaciones sobre representaciones internas en la que derivó el funcionalismo al tratar de dar cuenta de la naturaleza de las actitudes proposicionales (*cf*. Block 1996: 21, *supra*)– que nos ofrece Matthews (1989: 103):

> The R[epresentational] T[heory of] M[ind] holds the following: (a) propositional attitudes are relational, (b) among the relata are mental representations, (c) mental representations are symbols: they have both formal and semantic properties, and (d) mental representations have their causal roles in virtue of their formal properties.

Que esta singular costura de sintaxis y semántica se enhebra en la urdimbre de la idea de computación es subrayado con igual nitidez, *inter alia*, por Claplin (2002a: 13):

> In considering the nature of representation, it is important to recognize that at its very heart, the theory of effective computation relies on symbolic representations. The symbols stored on the Turing machine's tape are representations, and they explain and cause the behaviour of the Turing machine, and do so according to their semantic interpretation. […]
> [...] The computational trick was to construct representational vehicles whose physical, syntactic, and semantic properties converged. (Clapin 2002a: 13)

Pero –insistiría Fodor– que los símbolos causen la conducta *de acuerdo con* su interpretación semántica no entraña que sea su interpretación semántica *la causa de* su conducta: las propiedades sintácticas de los símbolos con las que esa interpretación semántica converge se ocupan de todo. La ruta que toma desde aquí el pensamiento de Fodor nos es ya conocida: estamos por fin en disposición de construir una teoría de los procesos mentales capaz de explicar las relaciones semánticas que consistentemente parecen darse entre aquellos pensamientos que se enlazan como causa y efecto, pero para hacerlo debemos asumir la existencia de símbolos internos, representaciones mentales dotadas de propiedades sintácticas –y también semánticas, insiste Fodor, aunque éstas, como hemos visto, quedan exánimes[306]. Debemos asumir, en suma, la existencia de una *lingua mentis*, puesto que:

> […] patently, there are going to have to be mental representations if this proposal is going to work. In computer design, causal role is brought into phase with content by exploiting parallelisms between the syntax of a symbol and its semantics. But that idea won't do the theory of mind any good unless there are *mental* symbols, mental particulars possessed of semantic *and syntactic* properties. There must be mental symbols because, in a nutshell, only symbols have syntax, and our best available theory of mental processes –indeed, the *only* available theory of mental processes that isn't *known* to be false– needs the picture of the mind as a syntax-driven machine. (Fodor 1985: 27)

Explicar, como se ha dicho, las relaciones semánticas que consistentemente parecen darse entre aquellos pensamientos que se enlazan como causa y efecto es –piensa Fodor– construir "[…] a theory of mental processes that succeed where

[306] Con el lema que Fodor (1975: 34) acuñó en su día, y que ha hecho fortuna incluso a la hora de cuestionarlo: "[…] no computation without representation" –o, en una expresión menos acendrada, "[…c]omputation presupposes a medium of computation: a representational system" (Fodor 1975: 27). *Cf*. Block (1980a: 28, 1996: 21, *supra*).

associationism (to say nothing of behaviorism) abjectly failed –a theory which explains how there could regularly be non-arbitrary content relations among causally related thoughts" (Fodor 1985: 27).

Los principios que de forma más o menos explícita rigen esta particular alquimia de la semántica en la sintaxis han sido desgranados con precisión por Horst:

> First, cognitive processes are *sequences* of intentional states. Now, according to [the] C[omputational] T[heory of] M[ind], to be in a particular intentional state is just to be in a particular functional relation to a mental representation. So if an organism is undergoing a cognitive process, it is passing through a sequence of functional relations to mental representations. Second, there are causal relationships between the intentional states that make up a cognitive process. […] Third, the causal connection between the states picked out is not merely incidental, but depends in a regular way upon the *syntactic properties* of the mental representations. […] Fourth, as in the case of a formal algorithm or a computer program, any semantic differences between mental representations are reflected by syntactic distinctions. (Horst 1996: 36)

Así que las ambiciones del cognitivismo no son –bien puede verse– de tono menor: de lo que se trata en suma es de alumbrar los vínculos entre lo que Sellars (1956), en su influyente *Empiricism and the Philosophy of Mind*, había dado en llamar el espacio lógico de las razones, por un lado, y el reino de la ley, o de la causalidad, por otro. No pocas veces, la altura de esa ambición ha conducido a una retórica que evoca casi a los *philosophes* del Siglo de las Luces y su determinación de convertirse en los "Newtons de la mente" (*cf*. Leahey 2005: 160):

> Nos interesa desarrollar un formalismo o "representación" en el que se describa [...] el conocimiento. Buscamos los "átomos" o "partículas" que lo componen, así como las "fuerzas" que actúan sobre él. (Winograd 1976: 9)[307].

Este audaz intento de engarzar el orden de la racionalidad y el orden de la causalidad es diseccionado con particular transparencia por Horst:

> In order for a sequence of representations to make up a rational, cogent train of thought, the question of *which* representation should occur in the sequence should be determined by the meanings of the earlier representations. In order for the sequence of representations to *make sense*, the later representations need to stand in appropriate *semantic* relationships to the earlier ones. But in order for a sequence of representations to be a *causal* sequence, the question of what representations will occur later in the sequence must be determined by the causal powers of the earlier representations. [...] But this can only be done if the semantic values of representations can be linked to, or coordinated with, the causal roles they can play in the production of other representations and the etiology of behavior (Horst 1996: 27)

[307] Declaraciones como ésta sirven a Dreyfus y Dreyfus (1988: 76) para calificar buena parte de los presupuestos metodológicos de la investigación en inteligencia artificial como "[…] a naive transfer […] of methods that have succeeded in the natural sciences".

Con aire mucho más desenfadado, la misma idea es bosquejada por el propio Fodor (1994b: 296), que atribuye su –digamos– descubrimiento a Alan Turing:

> What Turing did was to take the traditional analogy between minds and symbols absolutely seriously. Symbols have both semantic and material properties. [...] "Very well, then," Turing (more or less) said, "perhaps one could build a *symbol manipulating machine* whose changes of state are driven by the material properties of the symbols on which they operate (for example, by their weight, or their shape, or their electrical conductivity). And perhaps one could so arrange things that these state changes are semantically coherent [...]".

De este modo, la convergencia entre semántica y sintaxis acaba por perfilarse como un rasgo definitorio de la naturaleza de la representación mental. Así, claramente, parece suceder para Lycan (1986b), aunque la propia idea de convergencia quede elidida en una definición que orbita implícitamente en torno suyo:

> A representation is an inner state of an organism that has both a distinctive causal surface, responsible for the inferential relations [to other actual and possible representations], and semantical content, determined by the [causal and/or teleological] relations to things in the world. (Lycan 1986b: 161)

También Toribio (1995: 248) alude sólo de forma tácita a los goznes entre orden y significado: "[…u]na representación es […] una configuración física que tiene una lectura sintáctica y una lectura semántica". A las mismas palabras, en cambio, les seguía en Toribio (1991: 13) un somero esclarecimiento del modo en que ambas lecturas se incardinan de acuerdo con las tesis computacionalistas:

> […a]lthough computer processes are only sensitive to the syntax, the machine can be designed in such a way that the production of syntactic tokens makes sense given the semantic interpretation imposed by the problems that it is meant to solve.

El papel asignado a la computación en esa coyuntura entre forma y significado se vuelve igualmente perspicuo en el balance que ofrece Horst:

> […] machine computation is believed to provide answers to two questions: (1) How can semantic properties of symbols be linked to causal powers that allow the presence of one symbol s_1 at time t to be a partial cause of the tokening of a second symbol s_2 at time $t+\partial$? And (2) how can the laws governing the causal regularities also assure that the operations that generate new symbol tokens will "respect" the semantic relationships between the symbols, in the sense that the overall process will turn out to be, in a broad sense, rational?
>
> […] All in all, the computer paradigm shows that one can coordinate the semantic properties of representations with the causal roles they may play by encoding all semantic distinctions in syntax. (Horst 1996: 28-29)

La hipótesis de trabajo es, por consiguiente, que de la misma forma que es posible codificar sintácticamente toda diferencia semántica a la que puedan ser sensibles los

mecanismos computacionales de un autómata –precisamente de modo que *puedan* ser, o más bien parecer, sensibles a ellas–, toda diferencia semántica a la que los procesos cognitivos de humanos o animales puedan ser –o parecer– sensibles habrá de hallarse codificada en la sintaxis de un código interno: un lenguaje del pensamiento aún por descubrir, "[…] inmune" –como apunta Rivière (1991b: 145)– "a los contenidos y liberado de las impurezas de los factores semánticos y pragmáticos que en el pensamiento intervienen", no –cabría matizar– porque se los haya expulsado en una idealización por decreto, sino porque todas esas impurezas habrían quedado cristalizadas en el alambique de su sintaxis. O al menos –diría Fodor– todas las que realmente intervienen en nuestra vida mental desde una perspectiva propiamente psicológica: no así, sin duda, desde, por ejemplo, una perspectiva epistémica, en la que la verdad o falsedad de una creencia puede resultar capital[308]. Con las palabras del propio Fodor (1980a: 71), la clave de la cuestión es a su entender que "[…] truth, reference, and the rest of the semantic notions aren't psychological categories"[309].

El sustrato hilbertiano del funcionalismo ha sido recalcado reiteradamente y desde distintos ángulos en el transcurso de este estudio: con cierto detenimiento se ha venido bosquejando el modo en que la concepción funcionalista de lo mental que sirve de respaldo al cognitivismo arraiga en el programa de fundamentación formalista que Hilbert articulara para las matemáticas[310]. Hasta ese programa formalista es posible rastrear también, desde luego, la insistencia del funcionalismo, encabezada por Fodor (1980), en domeñar la contribución de la semántica de los estados mentales a la explicación psicológica encauzándola exclusivamente a través

[308] *Cf*. Burge (1986, *supra*) sobre la reformulación a la que el individualismo sometería al concepto de conducta para acomodar los objetivos explicativos de la psicología a los que hacen viable las restricciones que impone sobre la naturaleza de los estados mentales.

[309] A la pregunta de qué son entonces la verdad, la referencia y el resto de nociones semánticas responde Fodor en tono de burla: "What they are is: They are modes of *Dasein*. I don't know what *Dasein* is, but I'm pretty sure there's lots of it around […]" (Fodor 1980a: 71). Ésta es una de las pruebas de cargo que Dennett (1991b) emplea para caracterizar a Fodor como empedernido defensor de un mentalismo trasnochado –el que, a lo largo de las páginas de *La explicación psicológica* (Fodor 1968), quedara perfilado en su disputa contra el conductismo–, que, al ver amenazada por el propio avance de la investigación su caduca reivindicación del lenguaje psicológico, y habiéndose vetado las distintas variedades de criptodualismo a las que tantos otros recurren en parecidas circunstancias, se ve forzado a una retirada hacia el nihilismo.

Si bien parece desprenderse de él una lectura menos mordaz, igualmente distendido es el tono con que Devitt (1989: 392) se plantearía la cuestión de la importancia que su defensa de una psicología ceñida al contenido restringido de los estados mentales puede conceder a la referencia o el valor de verdad de dichos estados: "[…t]here is more to life than psychology" –felizmente, cabría añadir. Desde luego, que algo no sea importante para la psicología no conlleva que no sea importante *tout court*; cuando menos, desde luego, puede ser importante bajo una óptica epistemológica, o moral. Pensar si la psicología puede o no despojarse por entero de tales puntos de vista abriría, por supuesto, un debate tan difícil como necesario, pero que no nos es dado abordar aquí.

[310] Más allá de Hilbert, la destilación de la semántica en la sintaxis de la *lingua mentis* la enlaza estrechamente con el prístino sueño de la lengua perfecta cuyas reverberaciones hemos tanteado ya en Leibniz, Descartes y hasta Platón.

de las propiedades sintácticas de las representaciones internas. Al igual que el propio Fodor (1985: 26-27, *supra*), también Searle (1992: 203) aludiría un tanto a vuelapluma al desarrollo de la teoría de la prueba como uno de los pilares sobre los que –vanamente, a su juicio– se habría intentado sustentar la psicología cognitiva. En efecto, al reconstruir lo que él mismo denomina el relato primigenio del auge del funcionalismo, recala en el papel que el desarrollo de la teoría de la prueba pudo desempeñar a la hora de nutrir la credibilidad de la tesis –que Searle, por lo demás, considera un dislate– de que el significado de los estados internos de una mente artificial vendría dado *ipso facto* una vez que las relaciones sintácticas entre ellos estuvieran debidamente ajustadas. En la narración del propio Searle:

> "But what about the *semantics*? After all, programs are purely *syntactical*." Here another set of logico-mathematical results comes into play in the Primal Story. "The development of proof theory showed that within certain well known limits the semantic relations between propositions can be entirely mirrored by the syntactic relations between the sentences that express those propositions. Now *suppose* that mental contents in the head are expressed syntactically in the head, then all we would need to account for mental processes would be computational processes between the syntactical elements in the head. If we get the proof theory right the semantics will take care of itself; and that is what computers do: they implement the proof theory. (Searle 1992: 203)

Más explícitamente, Horst (1996) ha señalado que el precedente de Hilbert constituía la muestra más fehaciente de que la articulación de toda diferencia semántica bajo parámetros sintácticos era factible cuando menos en algunos ámbitos restringidos:

> Hilbert (1899), for example, demonstrated that it is possible to formulate a collection of syntactic types, axioms, and derivation-licencing rules that is rich enough to license as valid all of the geometric derivations one would wish for on semantic grounds while excluding as invalid any derivations that would be excluded on semantic grounds. (Horst 1996: 31)

Algo después recordaría también el propio Horst (2005) como el éxito del programa formalista de Hilbert a la hora de rescatar la fundamentación de las matemáticas del seísmo conceptual que había significado el desarrollo de geometrías no euclídeas se convertiría –lo hemos visto con algún detenimiento– en un paradigma cuyo ímpetu opera sin duda en la base del pensamiento funcionalista. No en vano, la búsqueda de una regimentación para la matemática ajena a las intuiciones espaciales que esas geometrías no euclídeas habían desacreditado era, precisamente, un apartamiento de –por así decir– la semántica de la geometría:

> [...t]he most influential strategy for formalization was that of Hilbert, who treated formalized reasoning as a "symbol game", in which the rules of derivation were expressed in terms of the syntactic (or perhaps better, non-semantic) properties of the symbols employed. (Horst 2005)

Así pues:

> The formalizability of limited symbolic domains show that semantic distinctions can be preserved syntactically and that the application of syntactic derivation rules can result in a semantically cogent sequence of representations. In crude terms, formalization shows us how to link semantics to syntax. (Horst 1996: 32)

La apuesta del cognitivismo, así vista, pasa por emular la proeza de Hilbert no ya únicamente para esos reducidos dominios lógicos o matemáticos, sino para la totalidad de la cognición humana: lo que ha llamado de Vega (1981) el metapostulado logicista de la psicología cognitiva se revela aquí en toda su profundidad. Pero en contraste con el entusiasmo que inicialmente había generado el proyecto, sus anhelados frutos resultaron ser menos asequibles de lo que parecía. El desencanto que Dreyfus (1972) supo ya retratar concienzuda e implacablemente queda perfilado en la concisa observación de Rivière (1991b: 148):

> […] los modelos sintácticos más ambiciosos y generalistas […] demostraron pronto que su ámbito explicativo era mucho más reducido que lo que se pretendió en un principio, y que sólo eran capaces de explicar la solución de problemas cerrados y de poca "densidad semántica", comportándose de forma aleatoria y muy poco "inteligente" en situaciones en que se aumentaba la carga semántica de los problemas.

El paulatino abandono de ese metapostulado logicista en el desarrollo histórico del cognitivismo ha sido, en efecto, lúcidamente relatado por Rivière (1991b), que en un brioso ejercicio de síntesis señala como sus hitos principales las investigaciones de Wason (1966, 1968) sobre las severas limitaciones de nuestra capacidad de razonamiento lógico y su dependencia respecto al contenido, luego reforzadas por los descubrimientos de Kahneman y Tversky (1973) y Tversky y Kahneman (1974) sobre sesgos y atajos heurísticos en los juicios que nos forjamos en condiciones de incertidumbre, junto con los trabajos sobre razonamiento analógico de Evans St. B.T. (1972), sobre los factores semánticos y pragmáticos involucrados, como mostraron Clark y Clark (1977), en la producción y la comprensión del lenguaje, y sobre la vaguedad inherente a la mayor parte de los conceptos que espontáneamente empleamos puesta de manifiesto por Rosch (1978), así como el desplazamiento de la idea de conocimiento basado en reglas hacia la de modelos mentales intrínsecamente semánticos, que tuvo su impulso decisivo en Johnson-Laird (1983). Una metáfora ontogenética sirve a Rivière para condensar los rasgos fundamentales de ese proceso:

> Al crecer, el sujeto de la psicología cognitiva –siguiendo, en cierto modo, un proceso contrario al que siguen los niños– se hizo cada vez menos serio, más difícil de formalizar, menos predictible y se alejó del modelo de racionalidad canónica que se le había marcado en un principio. (Rivière 1991b: 148)

Con la misma claridad, en efecto, con la que iba cobrando forma la primacía de cierta noción de lo sintáctico en la explicación de los procesos psicológicos, se iba

fraguando a la par, en el mismo seno del cognitivismo, la tesis de que a tal explicación le pertenece un lenguaje teórico propio, diferenciado tanto del empleado por los físicos como del de los fisiólogos, y uno de cuyos rasgos distintivos sería precisamente la carga semántica de los estados internos que se imbricaban en la explicación, su carácter rotundamente intencional[311]. No en vano asegura Pylyshyn (1984: 7, *supra*) que "[…] there are regularities and generalizations which can be captured using cognitive terms that could not be captured in descriptions using behavioral or physical (neurophysiological) terms" –la cifra de lo cognitivo es para Pylyshyn, ya sabemos, lo intencional. Aunque es indudable que el abandono del logicismo en el seno de la investigación experimental, tal como lo ha esquematizado Rivière (1991b), contribuiría a reforzar esa adhesión al valor explicativo de lo semántico, hay también motivos para pensar que constituye en realidad una fuerza inherente al funcionalismo y primigenia en su desarrollo, que, contrapuesta al énfasis formalista, forma parte de la complexión teorica de la psicología cognitiva.

También el propio Fodor reitera en muchos lugares su convicción de que, dicho en tono más desenfadado, "[…] we have no idea of how to explain ourselves to ourselves except in a vocabulary which is saturated with belief/desire psychology" (Fodor 1987: 15). De hecho, una segunda mirada a la manera en que Fodor plantea la condición de formalidad que a su juicio debe regir la explicación psicológica comienza ya a arrojar alguna sombra de duda sobre su coherencia interna. Así, recordemos, Fodor (1985: 26, *supra*) ha dejado dicho que "[…] to all intents and purposes, syntax reduces to shape" pero también, casi a renglón seguido, que "[…] there must be mental symbols because, in a nutshell, only symbols have syntax" (Fodor 1985: 27, *supra*). Algún ilegítimo deslizamiento ha de haberse dado en la argumentación cuando ésta nos permite concluir sin esfuerzo que nada que no sea un símbolo está dotado de forma, pues ninguna otra cosa –se nos dice– es la sintaxis y sólo los símbolos la poseen: así que ni los planetas ni las sillas, por mucho por que pueda parecerlo, tienen forma, no siendo símbolos. Dicho de otro modo: para que pueda ser cierto que sólo los símbolos tiene sintaxis ha de serlo igualmente que la sintaxis es después de todo algo más que mera forma, ya que mera forma, desde luego, no sólo los símbolos la ostentan. La imposibilidad de mantener a la par estos dos compromisos cruciales de la posición de Fodor sin vernos arrastrados al absurdo es un signo tan nítido como profundo de la ambivalencia respecto al valor de la semántica en la explicación psicológica que es endémica al cognitivismo, puesto que es precisamente la semántica –se argüirá *infra*– aquello que permite abstraer de entre las propiedades físicas de un símbolo aquellas que constituyen propiedades sintácticas, o, dicho de otro modo, aquello que habilita la atribución de unas propiedades sintácticas que, en efecto, sólo los símbolos poseen.

[311] Idéntica ambivalencia –no está de más apuntarlo– es fácil de revelar en el tratamiento que da el funcionalismo cognitivista –o, antes, y de forma quizá aun más patente, la cibernética, *cf*. *infra*– al papel de la teleología en la explicación de la conducta: el convencimiento de que se está logrando la traducción del vocabulario teleológico a un vocabulario mecanicista más sofisticado que el hasta entonces disponible ampara tanto la reivindicación de aquel como su vaciamiento.

No es infrecuente, de hecho, que en los propios círculos funcionalistas cierta aquiescencia respecto a la primacía de la sintaxis se acompañe de contundentes afirmaciones de la inelubilidad del vocabulario intencional, de acuerdo con las cuales la adscripción de contenido semántico a los estados internos de los organismos cuyo comportamiento tratamos de entender constituiría un trámite inexcusable –no, como suele decirse, un *mero* trámite. Así de tajante se muestra, por ejemplo, Cummins (1989: 135-136) –en un tono que evoca vivamente las reflexiones de Pylyshyn (1984, *supra*):

> The C[omputational] T[heory of] Cognition assumes that cognitive systems can be described by a set of content-ascribing generalizations that are autonomous in that they don't reduce to non-content ascribing generalizations.

Las más de las veces, no obstante, aun si se asigna a la semántica de los estados mentales –como condescendientemente– un papel en la explicación psicológica cuya importancia incluso se enfatiza con aire un tanto protocolario, la delimitación de dicho papel viene dada por una u otra versión de la condición de formalidad, sin que se reconozca francamente que el papel que se le está concediendo es en el mejor de los casos netamente pragmático, y en el peor enteramente superfluo. Podría apuntarse en estos casos, forzando el retruécano, que el papel en la explicación psicológica que la condición de formalidad concede a la semántica es, después de todo, meramente formal –o, como suele decirse, un mero *formalismo*. Esa salvaguarda mínima, casi resignada, de algún rescoldo exangüe del imperio del significado sobre nuestros pensamientos y acciones, es probablemente lo que tiene en mente el propio Cummins (1989: 130) cuando apunta que:

> Almost everyone is prepared to allow a causal role *of sorts* to content. A fairly widespread view seems to be that the states of a system that have content have causal powers, and that appeal to these causal powers is, or might be, central to psychological explanation. But it is generally held that the states in question have their causal powers not in virtue of having the contents they do, but in virtue of something else, e.g., their electrochemical potentials or their activation levels and connections.

Parece claro, a fin de cuentas, que si algo impulsa al funcionalismo a alejarse de la idea de primacía de la sintaxis y buscar cobijo en el contenido semántico de los estados mentales es el asedio del fisicalismo, y que ese trayecto fuerza también cierta renuncia al internismo: condición de formalidad y solipsismo metodológico, como ya se ha apuntado, resisten juntos o caen juntos. Contra la interpretación más extendida, sin embargo, van Gulick (1989) ha argumentado que las convicciones internistas acerca de cómo se debe caracterizar la semántica de las actitudes proposicionales –eludiendo precisamente las propiedades que, como la referencia o el valor de verdad, tenderíamos a considerar insobornablemente semánticas– divergen en realidad de la

propia teoría representacional de la mente que constituye el corazón del cognitivismo[312]. Con las palabras que inauguran su trabajo:

> The representational theory of mind is committed to the view that at least some mental states represent or refer to item in the outside world. Nonetheless, some proponents of the representational theory have argued that in our psychological theorizing, we should not individuate or taxonomize mental states in terms of the external items which they represent or to which they refer […]. They claim that mental states should be type-individuated entirely on the basis of their roles within the organism's internal causal structure and without reference to any external or contextual facts about the organism's environment. (van Gulick 1989: 151)

El objetivo de van Gulick es desmontar ciertos argumentos, que califica de "metafísicos", mediante los cuales se pretende establecer *a priori* la verdad del internismo apoyándose en la naturaleza local de la causalidad y en la naturaleza no causal de las propiedades semánticas. Se trata en realidad de tres argumentos que van Gulick engarza en uno general. El primero de ellos toma como premisas *(i)* que las explicaciones que factura la psicología cognitiva son explicaciones causales, y *(ii)* que las propiedades semánticas a las que alude el externismo (referenciales, veritativas) no son causalmente eficaces; de la primera premisa se deriva que *(iii)* los estados y procesos a los que aluden las explicaciones cognitivas deben venir taxonomizados en virtud de sus propiedades causales; de la segunda premisa, unida a la tercera, se concluye *(iv)* que la psicología cognitiva no debe taxonomizar los estados o procesos a los que apela en sus explicaciones según criterios semánticos externistas[313].

Es, como cabría esperar, el paso de *(i)* a *(iii)* lo que van Gulick cuestiona, argumentando –como ya hacía Burge (1986: 47-48, *supra*), *cf*. también Davidson (1963)– que del hecho de que una explicación sea de carácter causal no se sigue que las entidades teóricas a las que apela deban venir taxonomizadas en virtud de propiedades sometidas a leyes causales. "Detectar un depredador", por ejemplo, es un buen candidato para aparecer en una generalización de carácter causal, pero no sería un tipo de estado perceptivo taxonomizado por una propiedad que instancie una ley causal estricta: la propiedad de ser un estado perceptivo del tipo "detectar un depredador" ni siquiera es causalmente eficaz en sí misma, de acuerdo con la propia tesis de que la causalidad corresponde de hecho a las propiedades físico-químicas subyacentes, que van Gulick asume. Cualquier propiedad sintáctica, de hecho, sería de este tipo: si aparece en una explicación causal no es en virtud de su propia eficacia causal ni, *a fortiori*, de su sometimiento a leyes causales en sentido estricto, sino de la de los estados físicos en que se encarne.

---

[312] Si bien –conviene apuntar– su caracterización de ésta como, sin más ni más, la tesis de que algunos estados mentales representan *o se refieren* al mundo trivializa el representacionismo hasta convertirlo casi en una mera negación del idealismo.

[313] Se ha introducido, por mor de la claridad, una levísima modificación en la numeración de las premisas explicitadas por van Gulick: *(ii)* es la tercera de su lista y *(iii)* la segunda.

El segundo argumento retoma como premisa que *(iii)* los estados y procesos a los que aluden las explicaciones cognitivas deben venir taxonomizados en virtud de sus propiedades causales, y le añade *(v)* que los poderes causales de un estado psicológico no se ven afectados por cambios ambientales mientras tales cambios carezcan de repercusiones en la estructura intrínseca del organismo[314]; la conclusión que de ello se deriva es *(vi)* que la taxonomización de los estados mentales de cara a la explicación psicológica no debe contemplar hechos ambientales.

En su respuesta a este segundo argumento, van Gulick, de nuevo, sigue de cerca a Burge (1986): del mismo modo que el carácter causal de una explicación no implica taxonomización causal, una explicación coherente con el carácter local de la causalidad no exige taxonomización local. La elección de un criterio de taxonomización es en gran medida una cuestión pragmática, que puede atender o no a mecanismos causales y por tanto locales. Los "poderes causales" aludidos en *(iii)* bien pueden venir taxonomizados de forma no local, pero *(v)* sólo es verdadera si se refiere a poderes causales estrictamente locales: el paso de *(iii)* y *(v)* a la conclusión, *(vi)*, por tanto, descansa sobre un uso equívoco de "poderes causales". La respuesta de Fodor (1987: 41, 44) a Burge –a saber: que el carácter local de la causalidad implica superveniencia local de las propiedades causales, y por tanto obstruye toda taxonomización que no sea local– fracasaría según van Gulick precisamente por no reconocer el carácter pragmático, y por tanto sensible al contexto, de la taxonomización.

La debilidad del tercero de los argumentos *a priori* en pro de una concepción internista de la semántica de los estados mentales que van Gulick pretende desarbolar es, a su juicio, más patente. También la forma del argumento –que van Gulick atribuye a Fodor (1980a)– es más sencilla: de la premisa según la cual *(vii)* los procesos psicológicos, en la medida al menos en que sean procesos computacionales, carecen de acceso a hechos semánticos de naturaleza referencial o veritativa, se infiere que *(viii)* nuestra teoría de dichos procesos debería describir las representaciones sobre las que operan de suerte que se excluya la mención de tales hechos. Pero la premisa (vii), según la objeción de van Gulick, depende de la conclusión ilegítima del predicado "tiene acceso sólo a la sintaxis" a partir del

[314] Asumiendo –agrega van Gulick (1989: 152), precavido– que los cambios ambientales no conlleven alteración de las leyes naturales. Salvado el matiz, la formulación de la premisa no puede dejar de evocar las objeciones al antiindividualismo basadas en la idea de acción a distancia, a la que ya replicara Burge (1986: 50, *supra*). El propio van Gulick (1989: 154), en efecto, describe así la motivación de la premisa:

> How could an organism's causal powers be changed without some change in its intrinsic nature? To assume that an organism's causal powers could be changed by environmental changes that do not affect its intrinsic nature would seem to suppose some highly suspect form of action at a distance and to violate some very well entrenched scientific beliefs about how the natural world works.

predicado "tiene acceso a la semántica sólo a través de la sintaxis"[315] –que, de nuevo, van Gulick está dispuesto a conceder que se aplica a los procesos psicológicos:

> Facts about a representation's semantic content are facts about its relation to other (often distant) parts of the world, and thus cannot be directly recognized or directly acted upon by internal processes. Semantic properties *per se* cannot be immediate causal factors. A representation's semantic properties can influence the internal processes that operate on it only in so far as those properties are embodied in the representation's intrinsic causal structure. Internal processes can have *no direct immediate access* to semantic properties. However, it does not follow that internal processes have *no access at all* to the representations on which they operate. (van Gulick 1989: 158)

Lo que las observaciones aparejadas por van Gulick (1989) nos permiten establecer, no obstante, dista de ser que la renuncia a la semántica en la explicación psicológica, o su acotación a los parámetros de la noción de contenido restringido, esté injustificada o, menos aún, que debamos asumir un vocabulario psicológico plenamente semántico. Las conclusiones de van Gulick son, si se quiere, más humildes: que la polémica entre el internismo y el externismo no puede resolverse *a priori* –pues las presuntas pruebas *a priori* del internismo son endebles–, sino con argumentos que apelen a la práctica real de la teorización cognitiva.

La cuestión de los vínculos entre sintaxis y semántica se convierte entonces en una pregunta por los compromisos ontológicos a los que nos obliga el recurso a cierto vocabulario teórico: en la depurada síntesis que ofrece Horst (1996: 2), se trata al fin y al cabo de elucidar "[…w]hat are psychological theorists really committed to in their use of theoretical terms such as 'representation' or 'syntax'?". Cabe la respuesta de Fodor: la teorización psicológica reivindica la semántica de las creencias y los deseos –el único modo que podamos imaginar de entendernos a nosotros mismos–, si bien –o precisamente porque– condensa su eficacia causal, por medio de una regimentación sintáctica estrictamente formal, en propiedades físicas de los símbolos de la *lingua mentis* en que creencias y deseos se encarnan; sería un error, sin embargo, reducir las teorías psicológicas a teorías fisiológicas –o en última instancia físicas– acerca de la actividad nerviosa que conforma dichos símbolos. Cabe, también, la respuesta de Burge:

> I have seen no sound reason to believe that this use [of interpreted that-clauses and other intensional constructions –or what we might loosely call "intentional content"–] is merely heuristic, instrumentalistic, or second class in any other sense". (Burge 1986: 43).

Cabe, desde luego, preguntarse en qué medida la respuesta de Fodor se ajusta a la valoración de Burge: en qué medida el papel que Fodor concede al contenido intencional en la explicación psicológica no es *de hecho* meramente heurístico,

[315] Además, la réplica al segundo argumento bloquearía el paso de *(vii)* a *(viii)* incluso suponiendo que *(vii)* fuese verdadera.

instrumentalista, etc. –*de hecho*, porque las declaraciones del propio Fodor tanto parecen autorizar una conclusión como la contraria. En realidad, como apunta Horst (1996: 355), asumir que la ciencia ha de construirse con conceptos cuya delimitación sea capaz de atrapar regularidades causales y que al menos algunos de los conceptos medulares de la psicología, como los de creencia y deseo, vienen delimitados por propiedades semánticas, nos aboca o a la conclusión de que las propiedades semánticas son causalmente eficaces *per se* o la de que se hallan ligadas a otras que lo son[316] –de lo contrario, sólo quedaría, como augura el eliminacionista, tomar nota de la inconsistencia entre la estructura conceptual de la psicología y la que debería tener para conformar un discurso científico.

En este mismo ámbito fructifica la pregunta clave que, entre tantas otras, suscita la lectura de Rivière (1991b: 129), quien cifra lo que de novedoso tiene el cognitivismo en la utilización de un vocabulario intensional para explicar "[…] observaciones establecidas en términos extensionales" y en el engranaje de dicho vocabulario bajo nociones tomadas de la teoría de la computación, "[…] que implican el compromiso con un mecanicismo abstracto y formal": ¿en qué medida, entonces, la intensionalidad del vocabulario teórico subsiste a su regimentación formal; en qué medida queda subsumida por ésta y deviene instrumento heurístico o propedéutico?

Dicho de otro modo: es discutible que el voluntarioso abandono del logicismo que ha quedado bosquejado pueda culminarse –por ejemplo, restaurando la relevancia explicativa de la semántica de los estados mentales– sin quebrantar por ello compromisos nucleares del funcionalismo. En efecto: si, después de todo, la aparencia de que la semántica de los estados mentales concurre al comercio causal que liga estímulos y conductas es ilusoria, y viene propiciada por la impoluta concordancia de esa semántica con la sintaxis de las representaciones internas, tal vez no estemos forzados a admitir que la condición de formalidad pretenda una reducción de la semántica a sintaxis –pues no parece postularse que los hechos semánticos estén ontológicamente constituidos por hechos sintácticos: estarían reflejados o, si se prefiere, codificados en ellos–, pero sí cuando menos una reducción a la sintaxis del papel de la semántica en la explicación psicológica –una reducción estrictamente epistemológica–, mientras respecto a la relación ontológica entre propiedades semánticas y propiedades sintácticas parece guardarse un prudente silencio. O tal vez –como convincentemente argumenta Toribio (1991), y como Fodor (1989: 59) había constatado que, por razones a su juicio equivocadas, tiende a ocurrir– el paraje donde desembocaríamos al plegarnos a la condición de formalidad, si no matizamos nuestra concepción de la eficacia causal de lo mental, sea más bien el del epifenomenismo:

[316] En rigor debe anotarse que Horst (1996: 355) recoge una tercera posibilidad, a saber: que exista "[…] some interpreter that is sensitive to semantic properties and is the locus of the causal powers". Pero es difícil ver cómo leer esa tesis sin convertirla en una de las otras dos: o el intérprete es sensible a las propiedades semánticas y es el verdadero locus causal en el sentido de que refleja las propiedades semánticas sin que éstas tengan *sensu stricto* efectos sobre él, o el intérprete es sensible a las propiedades semánticas y este mismo hecho las hace causalmente eficaces.

> The path that takes us to mental epiphenomenalism is clear: 1) the causal powers of any event are completely determined by its physical properties; 2) although intentional properties supervene on physical properties, they can't be identified with them; 3) intentional properties, as intentional, are not causally responsible for behaviour, because they don't take part in the causal powers of the states to which they belong, *i.e.*, intentional properties are *epiphenomenal*. (Toribio 1991: 2) [317]

En un giro que evoca una vez más a Quine (1960: 280, *supra*), Heil (1989b), al igual que Toribio (1991), establece un estrecho nexo entre los vínculos de superveniencia que atarían a propiedades mentales y físicas –o semánticas y sintácticas, si se acepta tal reconstrucción– y un cuestionamiento del papel que pueda caber al significado en la explicación psicológica –que puede arrastrarnos incluso a conclusiones eliminacionistas que tilda de autorrefutatorias, *cf. supra*. El eslabón crucial es, una vez más, el papel de las propiedades semánticas en el control causal de la conducta, lo que provoca la impresión de que sean las observaciones antiindividualistas las que dan pábulo a la desconfianza sobre lo semántico o lo mental:

---

[317] Que una teoría enteramente sintáctica de la mente constituyera respecto de las propiedades semánticas de los estados mentales una suerte de epifenomenismo o de eliminacionismo, como apunta por ejemplo Devitt (1989: 369, *infra*), es una cuestión que queda abierta. La idea –vagamente entrevista en Fodor (1980a)– de que la semántica de los estados mentales –su referencia, su verdad– puede ser relevante en el ámbito de explicaciones ajenas a la psicología obraría, naturalmente, en contra de su eliminación, que defendió ardientemente Stich (1983) antes de desdecirse de buena parte de sus, otrora, conclusiones (Stich 1992, 1996).

Por otra parte, mostrar que una vez esquivado ese sumidero de la eliminación nos veríamos abocados a una posición epifenomenista y no a un velado paralelismo –como, *mutatis mutandis*, en Malebranche o Leibniz, o como en aquellos primeros tiempos de la psicología científica con los que a veces, por otros motivos, se ha vinculado al cognitivismo (*cf*. Rivière 1991b o Mandler 2002, *supra*) y cuyas veleidades paralelistas hemos visto ya que Watson (1913: 166, *supra*) repudiaba expresamente– pasaría por examinar si, de acuerdo con la especificación puramente sintáctica de los procesos psicológicos que imaginamos, cabría mantener que las relaciones entre las propiedades semánticas de dichos procesos se desenvuelven de acuerdo con sus propias leyes –la doctrina de la armonía preestablecida que en Leibniz cimienta la concepción paralelista de mente y cuerpo tendría su trasunto, desde este punto de vista, en el riguroso innatismo defendido por Fodor: algo debe asegurar *ab initio* la concordancia de mente y cuerpo, de razones y causas, de sintaxis y semántica–, o si más bien cada instanciación de alguna de esas propiedades semánticas no es más que el efecto de la instanciación de determinadas propiedades sintácticas, por así decir, subyacentes. Ciertas expresiones de Fodor –existe una "[…] armonía general entre las propiedades semánticas y las propiedades causales de los pensamientos" (Fodor 1985: 25), los roles causales de los estados mentales "[…] closely parallel the implicational structures of their propositional objects" (Fodor 1985: 23)–, desde luego, evocan vivamente el paralelismo, pero comoquiera que todo el peso de la explicación psicológica se hace recaer en las propiedades sintácticas sobre las que operan los procesos computacionales, la idea de leyes semánticas paralelas parece superflua, al menos a primera vista –lo cual ampara la interpretación epifenomenista de Toribio (1991), que arraiga en la de Fodor (1989) y Jackson y Pettit (1990), *cf. infra*. Está por ver, sin embargo, cómo podríamos articular sintaxis y semántica en otros ámbitos de explicación, en los que la verdad o la referencia de creencias y deseos no puedan descartarse. Éste es, en todo caso, un hilo de la argumentación que no cabe deshilvanar aquí.

> If something has a particular content, it does so not because of its intrinsic features, but in virtue of the way it –or rather the agent harboring it– is embedded in an environment. A pair of intelligent agents, physically indistinguishable down to the last molecule, might nevertheless be distinguishable intentionally if they are differently situated in the world. Differences that fail to be reflected in agent's interiors, however, seem behaviorally irrelevant. Features of agents in virtue of which they might be said to possess a given intentional property –a belief, say, or a want– should seem to play no role in the mechanism of action. If, for instance, the intentional content of a particular state is determined by its relations to external, possibly remote, states of affairs, that content might vary independently of its physiological (or syntactic) characteristics. How, then, could intentional content reasonably be expected to figure in rigorous explanations of intelligent behavior? (Heil 1989b: 348-349)

Pero, si bien es cierto que la sospecha recae sobre las propiedades intencionales de los estados mentales con mayor virulencia cuando éstas se entienden en un marco externista, ni mucho menos desaparecen en el entorno del internismo. Como se apronta a explicar el propio Heil, es en realidad la noción de superveniencia, y no su desarrollo externista o internista, lo que despierta los resquemores epifenomenistas, que nunca tardan en cobrar una lectura eliminacionista:

> Imagine […] that the relation between mental states possessing content and the biological states of intelligent creatures to whom those states belong is one of supervenience. […] We might suppose that the underlying states in question are internal, biological conditions of the creatures in question. Or we might think that the relevant physical "base" includes goings-on outside the creature. […]
>
> In either case, it is tempting to suppose that the emerging properties themselves do no *work*: causal powers they seem to possess in reality belong, exclusively to events occurring in the substrate. […] The supervenient items are, in that respect, epiphenomenal, nomologically inert: they *dangle*. (Heil 1989b: 354)[318]

La propuesta de Heil para esquivar esta conclusión parece requerir de una noción de sobredeterminación causal que acompañe a la de superveniencia. Cierto es que si lo mental superviene en lo físico, cualquier regularidad psicológica cuyo hallazgo pudiéramos atesorar, digamos $M_1 \rightarrow M_2$, podría más parsimoniosamente expresarse como $B_1 \rightarrow B_2$ si, en efecto, M_1 está ligado a B_1 por los lazos de la superveniencia, y hereda de B_1 –de acuerdo con Kim (1993a: 355, *supra*)– todos sus poderes causales, y si lo mismo, claro, puede decirse de M_2 y B_2. Aun así –se pregunta Heil (1989b: 355)–:

> […] what follows from this? Not, certainly, that the relation between M_1 and M_2 is not straightforwardly causal. To think otherwise would be to conflate an epistemological point about what we load into a particular lawlike generalization and an ontological point about causal relata. *B*-to-*M* do not reduce *M*s to *B*s, at least not in the sense of eliminating *M*s in favor of *B*s. Liquidity and solidity evidently depend on definite

[318] El uso de la expresión "nomological danglers" –algo así como *jirones* o *flecos nomológicos*– para referirse a eventos o propiedades que parecen no ajustarse a un conjunto dado de regularidades ha hecho fortuna desde que la emplearan H. Feigl (1958) y J.C.C. Smart (1959) refiriéndose a las propiedades fenoménicas de las sensaciones.

> underlying molecular structure. It does not follow from this, however, that liquidity and solidity do not or could not themselves have causal properties.[319]

Pero, dicho sea de paso, ese fruto no es el deseado: no nos basta con que los estados mentales existan, o puedan legítimamente establecer, en virtud de sus propiedades intencionales, las mismas relaciones causales que establecerían sólo en virtud de sus propiedades físicas como estados del sistema nervioso. A lo que aspirábamos era, más bien, a que nuestros estados mentales pudieran trabar, en virtud de sus propiedades intencionales, relaciones causales diferentes de las que propiciaran sus propiedades nerviosas; a que la intencionalidad instaurara alguna suerte de autonomía causal –de anomalía, si se quiere decir a la manera de Davidson– sobre la cual erguir –precisamente en el terreno de la epistemología– la autonomía de la explicación psicológica.

Un planteamiento externista agudiza la tentación epifenomenista, o eliminacionista –razona Heil– sólo en la medida en que aceptar relaciones causales entre las propiedades intencionales y la conducta pueda volverse cuestión más espinosa si a dichas propiedades intencionales se les asignan determinantes remotos[320], ya que esto parece contravenir la arraigada concepción mecánica, exclusivamente local, de la causalidad que venimos encontrando una y otra vez:

> Such a picture becomes still more compelling if we suppose that contentful mental states depend, not merely on goings-on *inside* agents, but also on agents' circumstances. Two creatures, biologically identical, might then harbour distinct thoughts. When we now try to imagine what it could be for mental states to operate as determinants of behavior, the situation appears hopeless. Any causal work done is done by local biochemical events in the brain. (Heil 1989b: 354)

A ojos de Heil, sin embargo, una semántica externista para los estados mentales no amenaza significativamente la relevancia del contenido intencional en la explicación de la conducta[321] –no más, al menos, de lo que pueda hacerlo ya el principio general de que sólo las propiedades físicas o químicas locales poseen genuinos poderes causales. Antes al contrario, se diría más bien que es en esa encardinación de la bioquímica nerviosa en los hechos del mundo donde anidan las perspectivas de una explicación psicológica autónoma. El argumento de Heil apela, aunque un tanto subrepticiamente, al carácter necesariamente histórico de la explicación de la conducta[322]:

---

[319] La cuestión de la eficacia causal de las propiedades disposicionales se ha examinado ya de la mano de Carnap (1932, 1938), Price (1953: 322), Geach (1957: 6), Armstrong (1968: 86) y Putnam (1975b), *supra*; y se retomará brevemente al hilo de los argumentos de Jackson y Pettit (1990: 205, *infra*).

[320] Se trata, por otra parte, de las mismas secuelas que deja el externismo –o cierta interpretación del externismo– sobre nuestra noción de autoridad de primera persona, secuelas que Heil (1989b) procurará también aliviar.

[321] Ni tampoco, de nuevo, su disponibilidad ante la introspección.

[322] Sobre el carácter ahistórico de las representaciones postuladas por la teoría computacional de la cognición, que contrasta con el claro componente histórico que comporta la adjudicación de contenido

> Explanations as to why an interior occurrence produces one sort of behavior rather than another may well depend on a certain relation the occurrence in question bears to external things […]. Suppose a particular neural event in the brain of a squirrel leads it to scurry up a tree –and not, say, to forage for nuts or run in circles. The event produces this response, perhaps, because this is the way the squirrel has come to be wired. But *why* is it so wired? In answering this question we may be obliged to recognize that the event in question bears a certain relation to some external item –the appearance of Spot, perhaps. In another creature […] a neural event with the very same relational properties may lead to utterly different sorts of response. (Heil 1989b: 355-356)

Resulta obvio que, al decir que el evento en cuestión mantiene cierta relación con determinado estímulo, Heil está hablando de una relación histórica. Puesto que la tarea es elucidar el modo en que el sistema nervioso del animal *ha adquirido* la particular hechura sobre la que reposa la conducta estudiada –en este caso, la conducta de huida ante un posible depredador–, sabemos de antemano que la explicación que construyamos será histórica. De qué conducta se trate es, en realidad, irrelevante: ya apele nuestra explicación a la filogénesis de la especie, en la medida en que la conducta sea innata, ya a la ontogénesis del organismo, en la medida en que sea aprendida, será, igualmente, una explicación histórica. Parece, en definitiva, que sólo manteniendo la explicación de la conducta en el ámbito interno, bajo los cánones del solipsismo metodológico, cabe preservar su condición ahistórica. Ensanchar el territorio de la explicación de la conducta para abarcar el ambiente en el que vive el organismo la enclava forzosamente bajo el dominio de la historia –que pudiera ser, tal vez, el territorio donde le aguarde su emancipación.

No es raro en todo caso –ya se ha apuntado *supra*– encontrar en el propio seno del funcionalismo gestos de aire conciliador, intentos, como el de Heil, de preservar a un tiempo –por así decir– tanto la primacía de la sintaxis como la eficacia de la semántica. Así, el reconocimiento de que la mente se guía por procesos sintácticos –es decir, neurofisiológicos, *cf*. Braddon-Mitchell y Jackson (1996: 257, *supra*)– puede ser declarado compatible con el de que venga también guiada por procesos semánticos –es decir, psicológicos: "The mind is indeed a syntactic engine. But that is consistent with its also being a semantic engine" (Braddon-Mitchell y Jackson 1996: 265). Con parecido espíritu aseguran Lepore y Loewer (1989: 166) que:

> […a]lthough Fodor endorses the formality condition, he also thinks that cognitive psychology contains true generalizations connecting propositional attitudes with each other, environmental conditions, and behavior. […] As Fodor emphasizes, the specification of propositional content in these generalizations is essential to their explanatory role. […] At first, this may seem incompatible with the claim that only formal properties of representations are relevant to the computations which produce behavior. However, there is no incompatibility as long as the contents of attitudes are specified in a way that respects the formality condition.

---

semántico a ciertas actitudes proposicionales en el discurso psicológico cotidiano, *cf*. Cummins (1989: 84, *supra*).

Ahora bien: está por ver que esta insistencia en la compatibilidad entre una regimentación sintáctica y una semántica, o entre una neurofisiológica y una psicológica, sea o no algo más que –según parece a primera vista– una declaración de buena voluntad. Un camino franco para valorarlo pasa por el examen de la respuesta que ofrecen los propios Braddon-Mitchell y Jackson a la pregunta por el papel que desempeña el contenido semántico de los estados psicológicos en la determinación de la conducta –*cf*. Braddon-Mitchell y Jackson (1996: 257, *supra*)– cuestión a la que dedican apenas la última página de su tratado. El problema cristaliza, a su juicio, en la legitimidad del uso explicativo de cláusulas adverbiales encabezadas por *qua*, empleadas para recalcar que un cierto estado mental desempeña éste o aquel papel en la explicación por razón de alguna de sus propiedades semánticas:

> Suppose that in Jones here and now, the kind of state that is her belief that a tiger is near is neurological state of kind *N*. *N* itself will simultaneously satisfy many descriptions, including '*N*', 'is belief that a tiger is nearby in Jones here and now', and so on. What makes it true that it is *N qua* belief with this content that explains Jones's running rather than, say, *N qua N*? (Braddon-Mitchell y Jackson 1996: 265)

Idéntica formulación puso ya sobre la mesa Fodor (1989: 59) al describir la preocupación –vana, a su entender– según la cual "[…] it is *not* compatible with physicalism that intentional states should be causally responsable for behavioral outcomes *qua intentional*", así como Toribio (1991), al preguntarse si el engranaje entre semántica y sintaxis que nos proporcionaría la idea de representaciones internas –sometidas, se entiende, a una regimentación computacional– conduce ineludiblemente al eliminacionismo acerca de las propiedades semánticas de lo mental, o acaso sobre lo mental *simpliciter* –es decir, si a fin de cuentas nos veremos forzados a renunciar a la convicción de que "[…] intentional states *qua* intentional, *i.e.* as having a particular meaning, are the ones responsable for our behaviour" (Toribio 1991: 2).

Se trata desde luego de una de las preguntas cardinales de, acaso, cualquier reflexión madura sobre la naturaleza de lo mental. Bajo el prisma de la distinción entre razones y causas, y por medio de un atinado ejemplo la planteaba también Dretske en un trabajo titulado, precisamente "The Explanatory Role of Content":

> Something possessing content, or having meaning, may *be* a cause without its possessing that content, or having that meaning, being relevant to its causal power. Shrieking obscenities may shatter glass but the semantics of these acoustic vibrations (their meaning) is irrelevant to their having this effect. […] Following Davidson, we can say that reasons are causes, but the problem is to understand how their *being* reasons, how their *having* semantic content, contributes to their causal efficacy. (Dretske 1985: 32)

Como suele ocurrir, la misma pregunta puede encontrarse ya en algún diálogo platónico: hay un bellísimo pasaje del *Fedón* en el que la distinción entre explicar una conducta a tenor de las razones que la impulsan –creencias, deseos, etc.– y hacerlo en

virtud de sus causas cobra tintes conmovedores. Dice Sócrates, relatando su decepción con las enseñanzas de Anaxágoras sobre el *Nous*:

> Me pareció que había algo muy parecido [en las explicaciones de Anaxágoras] a como si uno afirmara que Sócrates hace todo lo que hace con inteligencia y, luego, al intentar exponer las causas de lo que hago, dijera que ahora estoy aquí sentado […] porque mi cuerpo está formado por huesos y tendones, y que mis huesos son sólidos y tienen articulaciones que los separan unos de otros, y los tendones son capaces de contraerse y distenderse, y envuelven los huesos junto con la carne y la piel que los rodea. Así que al balancearse los huesos en sus propias coyunturas, los nervios al relajarse y tensarse a su modo hacen que yo sea ahora capaz de flexionar mis piernas, y ésa es la razón por la que estoy yo aquí sentado con las piernas dobladas. Y a la vez, respecto de que yo dialogue con vosotros diría otras causas por el estilo, aduciendo sonidos, soplos, voces y otras mil cosas semejantes, descuidando nombrar las verdaderas causas: que una vez que a los atenienses les pareció mejor condenarme a muerte, por eso también a mí me pareció mejor estar aquí sentado, y más justo aguardar y soportar la pena que me imponen. Porque, ¡por el perro!, según yo opino, hace ya tiempo que estos tendones y estos huesos estarían en Mégara o Beocia, arrastrados por la esperanza de lo mejor, si no hubiera creído que es más justo y más noble soportar la pena que la ciudad ordena, cualquiera que sea, antes que huir y desertar. (Fedón 98*c*-99*b*)

No termina de estar claro, tras tanto como ha llovido, que hayamos logrado desplegar una explicación del papel de nuestras creencias en la determinación de nuestras acciones menos vulnerable que la de Anaxágoras a las penetrantes objeciones socráticas.

Cadenas causales díscolas y leyes *cæteris paribus*

La escueta respuesta de Braddon-Mitchell y Jackson a su pregunta por el valor de las cláusulas *qua* en la explicación psicológica parecería tratar siquiera tímidamente de afianzar la tesis de que la semántica de los estados mentales –como las creencias de Sócrates acerca de lo mejor y más justo– es relevante en la fábrica de su comercio causal *en algún sentido* que desborda la eficacia de unas propiedades sintácticas que ellos –a diferencia de Fodor– identifican estrictamente, tipo por tipo, con sus propiedades neurofisiológicas –no, por tanto, huesos y tendones, soplos ni voces, sino actividad nerviosa. La idea básica es que una conducta queda adecuadamente explicada como efecto de –digamos– la conjunción de una creencia y un deseo *qua* creencia o deseo con tal o cual contenido si el papel que la creencia y el deseo desempeñan a la hora de causar la conducta –y que desempeñan, entiéndase bien, en virtud de sus propiedades sintácticas, neurofisiológicas, no de sus propiedades funcionales– forma parte de la definición funcional que nos autoriza a categorizar la creencia, o el deseo –es decir, tal o cual estado neurofisiológico–, precisamente como creencia o deseo con tal o cual contenido. Quizá regresar al ejemplo ayude a aligerar una argumentación fatigosa:

> When we causally explain Jones's running in terms of her believing that there is a tiger nearby, two conditions must be satisfied: the property that is her believing must do (some of) the causing, and it must do so as part of its functional brief as the property that is her believing. After all, if *N* did the causing of the running, but in some freakish way that had nothing to do with its belief role, it would be wrong to explain Jones's running away in terms of her belief. (Braddon-Mitchell y Jackson 1996: 265)

Podríamos imaginar, como se nos sugiere, que *N* satisface cierta definición funcional que lo categoriza como creencia de que hay un tigre cerca, pero sin embargo, dado que Jones desea acercarse a un tigre, el contenido de su creencia no contribuye a causar su conducta de huida: ésta se debe en realidad a que *N* (la creencia de que hay un tigre cerca) causa de modo azaroso y ajeno a la definición funcional que lo constituye como tal creencia –ajeno a su "función normal" en Jones, dicen Braddon-Mitchell y Jackson (1996: 266)– "una serie de espasmos musculares" que hacen que Jones corra para obtener alivio, casualmente en dirección contraria al tigre. Desde luego, no sería apropiado decir en tal caso que la creencia de que hay un tigre cerca hubiera causado la conducta de Jones *en tanto que creencia de que hay un tigre cerca*.

El planteamiento de Braddon-Mitchell y Jackson entronca así con la discusión, ya clásica en el ámbito de la teoría de la acción, acerca de la importancia de atender a posibles cadenas causales anormales –desviadas, díscolas– a la hora de determinar la voluntariedad de una conducta, o de entender el papel de creencias y deseos en la génesis de la conducta en general. El problema se remonta a un ingenioso ejemplo con el que Chisholm trataba de desacreditar cualquier reconstrucción de la noción de propósito en términos de nociones no teleológicas de causalidad:

> […] cierto sujeto desea heredar una fortuna; […] cree que si mata a su tío heredará una fortuna; […] esta creencia y este deseo le excitan de tal modo que conduce con excesiva rapidez, con el resultado de que, accidentalmente, atropella y mata a un peatón que, sin que el sobrino lo sepa, no es otro que su tío. (Chisholm 1966: 29-30)

No es difícil apreciar el parentesco entre las preocupaciones de Chisholm y el análisis de Braddon-Mitchell y Jackson. Ya Goldman (1970: 62) matizaba el trabajo de Chisholm haciendo ver que lo que nos falta por elucidar es la "forma característica" en la que nuestros deseos y creencias deben desembocar en nuestras acciones si hemos de calificar dichas acciones como intencionadas o voluntarias; en la misma línea perseveraría Davidson (1980), aunque sólo para darse finalmente por vencido y conceder su incapacidad de definir en qué consista tal forma característica. Parece claro que la propuesta de Braddon-Mitchell y Jackson se cifra en considerar que una actitud proposicional debe figurar en la explicación causal de una conducta como tal actitud proposicional cuando el estado neurofisiológico *N* al que es idéntica figura entre las causas efectivas de la conducta en cuestión, y lo hace de la manera característica –digamos, normal: no desviada, díscola, ni "freakish" (Braddon-Mitchell y Jackson 1996: 265, *supra*). Como era de esperar, la vaga idea de cadena causal característica se reconstruye ahora apelando a la propia definición funcional

que permite categorizar a *N* como tal o cual actitud proposicional: una cadena causal es característica si forma parte de de ese "functional brief" (Braddon-Mitchell y Jackson 1996: 265, *supra*). Pero las cosas pueden verse al revés: cabría argumentar que el problema de las cadenas causales desviadas no infecta sólo a la teoría de la acción, sino también a la lectura funcionalista de los conceptos básicos de la psicología cotidiana. Es decir, que el proyecto de labrar análisis funcionales exhaustivos de nociones como la creencia o deseo tal como las maneja el sentido común se ve estorbado, si no irremediablemente obstruido, por la existencia de cadenas causales desviadas que, si bien el sentido común parece intuitivamente reconocer con cierta facilidad como tales, no resulta tan fácil aislar con los recursos del análisis funcional. La objeción no es, desde luego, definitiva, pero sí onerosa para Braddon-Mitchell y Jackson, sobre quienes recae la carga de dar con una solución satisfactoria al problema de las cadenas causales desviadas, o bien con una explicación convincente de por qué dicho problema no afecta a su concepción de la relevancia explicativa propia de lo mental.

Quizá las dificultades que afrontan Braddon-Mitchell y Jackson sean después de todo una secuela natural de su afán por preservar la relevancia explicativa de los conceptos psicológicos a la par que deniegan a los estados mentales a los que dichos conceptos aluden toda eficacia causal que no sea la de los estados neurofisiológicos a los que a su juicio son estrictamente idénticos –caso por caso y tipo por tipo. Lo que se nos sugiere es, en suma, que la regimentación semántica, o psicológica, sólo es consistente con la sintáctica, o neurofisiológica, en el sentido de que hay una descripción parcial, incompleta, incierta y confusa de la realidad que apela a propiedades semánticas, o psicológicas, y que no contradice la descripción total, completa, firme y clara que proporciona la apelación a propiedades sintácticas, o neurofisiológicas. No se puede negar aquí –como en el caso de la presunta autonomía de la explicación psicológica– que estemos ante un sentido cabal de la idea de consistencia entre dos explicaciones. Pero tampoco se puede obviar que resulta un tanto engañoso otorgar a la mente el carácter de mecanismo guiado por procesos semánticos cuando todo lo que estamos dispuestos a conceder es que los procesos sintácticos que de hecho la guían se pueden describir mal que bien en términos semánticos, pero en ningún caso que los procesos semánticos en cuestión puedan imponerse sobre los sintácticos y ejercer una eficacia causal propia.

Acaso más prometedor sea el intento de Cummins (1989) de respaldar la irreductibilidad de la semántica de los estados mentales –es decir, de construir una noción de su eficacia causal capaz de hacer irrenunciable su valor explicativo– empleando para ello las mismas herramientas de las que el funcionalismo se había servido, de la mano de Putnam (1960, 1963a, 1967a, 1967b, *supra*), para articular una concepción de las relaciones entre mente, cerebro y conducta ajena tanto a la tesis de identidad psicofísica como al conductismo lógico –fundalmentalmente, la tesis de realizabilidad múltiple y la distinción entre enunciados de identidad con alcance de tipos y con alcance de casos–, así como de su ulterior despliegue en las nociones de vocabulario teórico y captura de generalizaciones. Procede, por motivos que hemos

revisado minuciosamente, convenir al menos con carácter provisional en el punto de partida fijado por Cummins: que lo interesante a la hora de disipar las dudas respecto a "[…] si el contenido [semántico] es una causa" no sería averiguar si "[…t]here are true singular content-ascribing causal statements in which a cause (or causal factor) is identified via its content" –ya sabemos que incluso concediendo la condición de formalidad esto bien puede ser cierto, pues esa causa que *identificamos* en virtud de su semántica puede sin embargo *ser* una causa en virtud de su sintaxis, a la que su semántica se pliegue sin borrón–, sino decidir si "[…i]n a true singular content-ascribing causal statement, having the content ascribed is (at least sometimes) the 'causally relevant factor'" (Cummins 1989: 130) –de modo que el refugio formalista que acabamos de ensayar quede expresamente vedado. Así las cosas, el razonamiento que justificaría la condición de formalidad de Fodor –y que, quizá en menor medida que las consideraciones solipsistas avivadas por la intensionalidad del lenguaje mentalista (*supra*), también habría contribuido a inspirarla– puede reconstruirse así: "[…] when we come to generalize singular content-ascribing causal statements, we find that the lines of nonaccidental generalization are traced out, not by content, but by such things as electronic properties or patterns of activation or abstract shape" (Cummins 1989: 131).

Ahora bien, si las generalizaciones en las que pudiéramos articular esos enunciados causales singulares en los que se atribuye un contenido semántico a un estado mental resultaran extenderse "[…] along lines traced by content" (Cummins 1989: 132) –o, a la inversa, si formular esas generalizaciones en términos frugalmente sintácticos se revelara como un ejercicio repleto de disyunciones arbitrarias y potencialmente infinitas, como el propio Fodor ha argüido insistentemente que ocurriría si tratáramos de formularlas en un vocabulario neurofisiológico[323]–, entonces las mismas razones que nos impelen a aceptar la condición de formalidad nos obligarían a rechazarla. Pero es entonces cuando el estribillo ya familiar de los argumentos funcionalistas comienza a adivinarse:

> Systems that instantiate the same function –all adding machines, for example– have something in common that can be captured only by generalizing over semantic properties […] There is nothing special about cognitive functions in this connection, and hence the causal status of content rests on no special thesis about cognition held by the C[omputational] T[heory of] C[ognition]. (Cummins 1989: 133)

Desde luego, Cummins despliega esta aplicación de una tesis tan medular al funcionalismo como la de realizabilidad múltiple de manera mucho más cuidadosa. Se necesitan dos premisas, a su entender, para que de las propias tesis cognitivistas se siga como conclusión que la semántica de los estados mentales cobija una eficacia causal propia, que la sintaxis no podría usurpar. La primera de estas dos premisas es que si el cognitivismo es cierto, existen generalizaciones que involucran la

[323] *Cf.*, por ejemplo, Fodor (1997: 156, *supra*), en relación con la idea de *gerrymandering*.

adscripción de contenido semántico –algo que la condición de formalidad, como hemos visto, garantiza. Más en detalle:

> A cognitive capacity is specified, according to the C[omputational] T[heory of] C[ognition], by a cognitive function, i.e., by a system of generalizations that are or approximate epistemic rules. Since epistemic rules are generalizations defined over things with truth values, it follows […] that the generalizations in virtue of which something counts as cognitive, are going to characterize that thing in terms of content. Thus, if there are cognitive systems as envisaged by the C[omputational] T[heory of] C[ognition], there are generalizations involving content ascriptions. (Cummins 1989: 132)

La segunda premisa, en cambio, es polémica: no es posible –sostiene Cummins– purgar las adscripciones de contenido semántico que intervienen en dichas generalizaciones –es decir, reducir las generalizaciones a otras en las que el contenido semántico haya desaparecido–, porque lo impide la realizabilidad múltiple de los contenidos semánticos en estructuras sintácticas, que es tan de sentido común como pueda serlo la de lo mental en lo físico.

> […] these generalizations involve content ascription essentially; they are not going to reduce to (be replaceable by) non-content-ascribing generalizations because things with different electronic properties or abstract shapes or patterns of activation can satisfy the same content-ascribing generalizations. (Cummins 1989: 132-133)

Así que –concluye Cummins (1989: 133)– si la eficacia causal se cifra en la capacidad de develar generalizaciones nómicas, la verdad del cognitivismo –" the existence of cognitive systems as envisaged by the C[omputational] T[heory of] C[ognition]"– conlleva, en conjunción con la posibilidad de formular generalizaciones verdaderas en las que se atribuye contenido semántico a los estados internos de un sistema cognitivo y con el hecho presuntamente evidente de que distintas sintaxis pueden subyacer a la misma semántica, que la semántica de los estados mentales es causalmente autónoma y, con ello, que la condición de formalidad no sólo debe impugnarse, sino que su rechazo resulta, después de todo, en cierto sentido consustancial al cognitivismo[324]. En el fondo, "[…i]t is tempting to suppose that no

[324] Analizar la cuestión en términos de enunciados contrafácticos en lugar de hacerlo en términos de generalizaciones conduce, de acuerdo con Cummins (1989: 134-136), a idénticas conclusiones. Si bien es cierto a su juicio que no nos será dado hallar enunciados singulares que identifiquen un estado psicológico –o, en general, un estado interno de un sistema cognitivo– en virtud de su contenido semántico para los que puedan fijarse consecuencias contrafácticas que no se deriven también de algún enunciado singular que identifique dicho estado interno en virtud de propiedades semánticas, pero si atendemos a enunciados contrafácticos acerca de tipos de estados internos clasificados en virtud de propiedades semánticas –y no a enunciados singulares–, las mismas razones que nos llevan a admitir la existencia de generalizaciones semánticas inexpresables en términos sintácticos nos abocarán también a hacer lo propio para el caso de generalizaciones semánticas de orden contrafáctico. El análisis en términos de contrafacticidad, en suma, nos reconduce al análisis en términos de generalizabilidad.

theory of cognition can afford to ignore intentionality, since cognition is essentially an intentional phenomenon" (Cummins 1989: 139).

La principal dificultad que reviste el argumento de Cummins tiene que ver con el traslado del argumento de realizabilidad múltiple desde la relación entre estados psicológicos y estados neurológicos hasta la relación entre la semántica de los estados psicológicos y su sintaxis. En la versión clásica de ese argumento –que hemos analizado relativamente a fondo– desempeña un papel crucial el hecho de que los criterios taxonómicos que ofrece el lenguaje de la psicología se toman como una abstracción sobre los que ofrece el lenguaje de la neurología; de ahí que determinadas generalizaciones formuladas en términos psicológicos puedan carecer de reflejo –se argumenta– en términos neurológicos. Desde luego, las propiedades neurológicas pueden tomarse también como abstracciones –sobre propiedades electroquímicas, supongamos. Así ocurre también con las propiedades sintácticas, que –lo hemos visto– se describen canónicamente en el seno del cognitivismo como abstracciones sobre propiedades físicas. Sin embargo, que la descripción semántica de un estado mental conlleve un nivel de abstracción más elevado que su descripción sintáctica es controvertido, puesto que lo que se estipula es precisamente que, en el desentrañamiento de la *lingua mentis*, cualquier diferencia semántica habrá de quedar vertida en una de orden sintáctico. Sintaxis y semántica, desde este punto de vista, se encontrarían en el mismo nivel de abstracción –o eso cabría argumentar en pro de la condición de formalidad–, mientras que neurología y psicología estarían en niveles desparejos. El resultado es que en defensa de la primacía de la sintaxis siempre podrá alegarse que las propiedades sintácticas de las que hablamos están articuladas precisamente para recoger esas generalizaciones semánticas que el argumento de Cummins concluye que se le escaparían[325]. Pero ese –claro está– es el mismo lugar del que partimos, y no sería raro llegar a él con la frustrante impresión de haber caminado en círculo.

También Toribio (1991) intenta reconstruir la idea de eficacia causal de lo mental de suerte que la articulación computacional de sintaxis y semántica no acarre el desahucio del significado. El punto de partida de su análisis es una nítida distinción –que venimos ensayando reiteradamente de la mano de Pylyshyn (1984)[326]– entre el hecho de que eventos particulares –o sea, instanciaciones concretas de propiedades concretas– constituyan causas de otros eventos particulares, y el hecho de que ciertas clases de eventos muestren determinadas regularidades respecto a las clases de eventos que causan o los causan, regularidades que expresamos en forma de leyes. La observación elemental de que caben siempre diversos criterios para la formación de clases sirve entonces para comenzar a

[325] En realidad, como hemos visto en el transcurso de la discusión en torno a la idea de realizabilidad múltiple (*supra*), estrategias argumentativas muy similares se han empleado también en el intento de desarbolar la interpretación antireduccionista de observaciones como las de Putnam (1967a): *cf*. por ejemplo Bechtel y Mundale (1999), Shapiro (2000, 2004).

[326] También –recuérdese– de la de von Wright (1971: 43, *supra*).

entreabrir la puerta a que propiedades de distinta índole puedan atesorar alguna eficacia causal:

> Only individual events can be causes. But, at the same time, the necessary character of the regularities expressed by a causal law depends on such regularities being established not between particular events, but between *types* of events. Now, since particular events can be referred to by many different expressions, some of which don't mirror any of the properties that turn them into causes or effects of other events, the criteria for grouping together particular events into events of the same type –events of the type that can appear in a causal law– must only focus on those properties that can be shown to be causally efficacious. (Toribio 1991: 5)

Así, que una determinada propiedad *P* se halle revestida de la eficacia causal que se discute a lo mental o a lo semántico, "[…] comes down to the question of whether or not there are causal laws involving *P*" (Toribio 1991: 5). En suma:

> […] a property is causally efficacious if it is a property in virtue of which the objects that instantiate it can be subsumed by laws, possibly *ceteris paribus* laws. (Toribio 1991: 12)

Aunque pueda parecer en este punto que el razonamiento de Toribio se acerca a las conclusiones que ya han quedado trazadas acerca de la diferencia entre la noción paramecánica de causalidad que trasluce en la idea de poderes causales –como en Kim (1993a: 355) o en Bickle (2006: §2.7), *supra*– y una concepción de la causalidad bajo el prisma de la regularidad nómica –que hemos adjudicado a Davidson–[327], lo cierto es que si nos preguntamos qué dota a una ley en la que se expresa una cierte regularidad del marchamo de ley causal, la respuesta esbozada por Toribio es, en principio, enteramente afín a la idea de que los genuinos poderes causales operan únicamente en el ámbito de lo más pequeño, de los constituyentes últimos de la materia –si bien lo hacen más o menos del mismo modo, parece, que en la mecánica de los cuerpos macroscópicos. Se nos aclara, en efecto, que:

> […] what makes a causal law a proper law, and not a mere regularity with enough statistical support, is the existence of a micro-physical mechanism or structure that is shared by the different macro-types of events that are subsumed by that law. (Toribio 1995: 5)

Asumida esa idea de la causalidad, quedan aparejadas las firmes ligaduras que han de trabar a cualquier propiedad a la que pretendamos atribuir algún efecto con otras que pertenezcan a ese dominio privilegiado donde, como decía Bickle (2006: §2.7, *supra*), "[…] the rubber meets the road". Sin embargo, cabe anticipar, no es difícil ver cómo la idea de realizabilidad múltiple se infiltra ya entre esas ligaduras, como para destensarlas:

---

[327] Y a pesar de que Toribio (1991: 23), de hecho, cita al propio Davidson como fuente de inspiración de su concepción de la causalidad.

> In other words, a macro-type of events can be considered causally efficacious only if it supervenes on micro-physical events –*perhaps different on different occasions*– in such a way that the causal powers of the former are explained by the causal powers of the latter. (Toribio 1991: 5, énfasis añadido)

La eficacia causal de una propiedad se cifra, entonces, en el entramado nomológico en que se imbrique, y es la propia noción de ley lo que la traba a la existencia de mecanismos causales con los que dicha propiedad guarde una determinada relación –una relación que es controvertido identificar con la de superveniencia, toda vez que se afirma que dichos mecanismos pueden ser distintos en distintas ocasiones sin que, al parecer, varíe la propiedad en cuestión, o al menos el tipo de sucesos que queda delimitado por ella. Sea como fuere, en la medida en que existan leyes –aunque sea leyes *cæteris paribus*– que expresen regularidades en las que se hallen involucradas propiedades semánticas, no habrá inconveniente en concluir que:

> […t]he semantic properties of mental states can be seen as causally efficacious with respect to different behaviours, although the equivalent class to which they belong is not describable in terms of features that are projectable in a physical vocabulary. Of course there must be a physical description, as there has to be a physical implementation of the properties responsible for the semantic causation, but that physical characterization is not the relevant description for the explanation of their causal efficacy […]. (Toribio 1991: 14)

Tenemos entonces, por una parte, que los poderes causales de las clases de eventos definidas en virtud de propiedades macrofísicas *quedan explicados* por los poderes causales de los eventos microfísicos particulares en los que cada elemento de dicha clase superviene (Toribio 1991: 5, *supra*), pero, por otra, que la descripción física *no es relevante para la explicación* de los poderes causales de las propiedades semánticas de los estados mentales –que son, se sobreentiende, precisamente el tipo de propiedades que pueden definir clases de eventos como las mencionadas.

La apariencia de contradicción es nítida: una cosa queda explicada por otra, pero ésta no es relevante para la explicación de aquélla. Las rutas abiertas para disolver tal contradicción también aparecen despejadas: visto que *B* queda explicado por *A* pero *A* no es relevante para la explicación de *B*, o bien aquello por lo que *B* queda explicado y aquello que no es relevante para la explicación de *B* no son exactamente lo mismo –es decir, "*A*" es ambiguo–, o bien no son exactamente lo mismo aquello que queda explicado por *A* y aquello para cuya explicación *A* no es relevante –o sea, que la ambigüedad reside en "*B*"–, o, por último, a lo que se llama explicación en uno y otro caso no es exactamente a lo mismo –se usa ambiguamente, en suma, el término "explicación". Si el equívoco no se halla en los *relata* –dicho de otro modo–, tal vez resida en la propia relación. De no ser así –de no mediar ambigüedad alguna–, el territorio en que estas consideraciones desembocarían no podría sino resultarnos conocido: las clases de estados de un organismo o un autómata que podamos delimitar mediante el vocabulario psicológico pueden resultar *explicativamente relevantes* en la medida en que acogen generalizaciones que

de otra manera habría que formular recurriendo a poco airosas conjunciones o disyunciones de clases de estados punteadas en un vocabulario estrictamente físico, pero no cabe decir, sin embargo, que resulten *causalmente eficaces* en el pleno sentido de la expresión en que lo son las propiedades físicas –es la conclusión de Jackson y Pettit (1990), refrendada como acabamos de ver en Braddon-Mitchell y Jackson (1996).

Ante lo que nos encontramos, tal como evocadoramente lo describen Jackson y Pettit (1990) es con el trasunto de un viejo problema que aqueja a la concepción cartesiana de la mente –el argumento de la sombra de la fisiología, dicen Jackson y Pettit–, y que parece abocarla a reemplazar el interaccionismo por un resignado epifenomenismo[328].

> It is observed that it is very plausible that in principle a complete explanation of each and every bodily movement of a person can be given in terms of their internal physiology, with their neurophysiology playing a particularly important role, along with interactions of a physical kind between their physiological states and their environment. There are no mysterious, unclosable-in-principle gaps in the story medical science tells about what makes a person's arm go up. The conclusion then is that the sort of properties that feature in the dualist story are causally irrelevant to behaviour; and we are led to the familiar objection to dualism that the interactionist variety of dualism has to give way to an epiphenomenalist variety –and so much the worse for dualism! (Jackson y Pettit 1990: 198)

Pero una réplica casi punto por punto del argumento –mantienen Jackson y Pettit– pone también en duda la eficacia causal de las propiedades psicológicas y, en particular, de su vertiente semántica, "[…] despite the fact that functionalism is compatible with a purely materialistic view of the mind":

> Take, for instance, content and our commonsense conviction that content is causally relevant to behaviour, our conviction that the fact that a certain state of mine has the property of being the belief *that p* or of being the desire *that q* is causally relevant to my arm moving in a certain way. […] How can that be, given the just discussed fact that a complete explanation in principle entirely in physiological terms of my behaviour is possible? For the kind of property content is identified with in the functionalist story will not appear anywhere in that story. What will matter at the various points in that story will be the physiological, and particularly neurophysiological, properties involved, whereas, as has so often been emphasised, what matters from the functionalist perspective for being a certain kind of mental state is not the nature of the state neurophysiologically speaking but rather the functional role occupied by the state. One way of putting the point is by saying that what drives behaviour is the physiological

[328] O tal vez, a la luz del modo en que las propiedades semánticas de los estados mentales se ven truncadas para poder ceñirse a sus propiedades sintácticas, a una nueva variedad de paralelismo: *cf.* por ejemplo, Fodor (1985: 25), *supra*.

Tanto Jackson y Pettit como, aunque sólo para combatirla, el propio Fodor (1989: 59, *infra*) coinciden, no obstante, en la interpretación epifenomenista que hemos encontrado ya en Toribio (1991: 2, *supra*), y que contrasta con la lectura eliminacionista que Devitt (1989: 369, *infra*) tratará de desmantelar.

> nature of the various states, not the functional roles they fill. How then can functional role, and so content according to functionalism, be a causally relevant property? (Jackson y Pettit 1990: 198)

Bajo el influjo de estas consideraciones –apuntan luego Jackson y Pettit (1990: 206)–, "[…t]he intuition that functionalist accounts of content make content epiphenomenal is a strong one": se trata tal vez de la intuición crucial en la etiología de lo que Fodor (1989: 59) había descrito como "[…a]n outbreak of epiphobia ([…] the fear that one is turning into an epiphenomenalist) […]", que habría proliferado ya por entonces en el ámbito de la filosofía de la mente, y que el propio Fodor se propondría aliviar. Sin embargo, una reivindicación del papel explicativo de la semántica de los estados mentales es para Jackson y Pettit, si cabe, más apremiante que para el propio Fodor, habida cuenta de que a su juicio los estados mentales no vienen taxonomizados, como piensa Fodor, a tenor de los descubrimientos que la psicología científica pueda ir costrastando al respecto, sino más bien de lo que el sentido común nos dice sobre ellos[329]. En el contexto del discurso cotidiano en torno al modo en que nuestras creencias o deseos causan nuestra conducta, desde luego, es innegable que "[…] the properties we are invoking must be the known or guessed about functional roles, not the unknown nature of the occupiers of those roles" (Jackson y Pettit 1990: 200)[330]. Esto, al mismo tiempo, frustra la táctica más expeditiva a la que Jackson y Pettit –en la convicción de que la concepción funcionalista de lo mental, mediante el mero trámite de relativizar la identidad postulada a especies, sujetos o instantes, puede convivir sin fricciones con una tesis de identidad psicofísica entendida con alcance de tipos (*cf. supra*)– podrían recurrir para deshacerse del problema: insistir en que la eficacia causal de un estado mental –o, si se quiere, de su semántica– es estrictamente la misma que la del estado físico con el que se identifica.

Otra forma aparentemente sencilla –pero baldía, apuntan Jackson y Pettit (1990: 202)– de zanjar la cuestión sería añadir a la tesis de que los estados mentales supervienen en estados físicos un principio según el cual si la base de superveniencia de una propiedad dada es causalmente eficaz con respecto a un determinado efecto, entonces también lo es la propiedad que superviene en ella; unido al hecho de que la

---

[329] *Cf. supra*: también la amenaza de trivialidad se cierne, como vimos, más gravemente sobre el funcionalismo analítico en virtud de su compromiso con el conocimiento de sentido común.

[330] El ejemplo que aportan Jackson y Pettit (1990: 200) acaso aclare la importancia que desde su óptica, tan estrechamente ligada a la explicación psicológica ordinaria, reviste consagrar la eficacia causal de lo mental –o al menos, como de hecho ocurrirá, su relevancia explicativa:

> When I explain your behaviour by citing your belief that it is about to rain, I am surely explaining your behaviour in terms of something I know about you, or at least that I think that I know about you. I am not saying that there is some internally realized property, I know not what, which is causally relevant to your behaviour. […] I am rather explaining your behaviour in terms of something I know about you; and as I do not know, and know that I do not know, about the nature of your internal physiological states, it can only be the relevant functional role which I am citing as the property which you instantiate which is causally relevant to your behaviour.

eficacia causal, de cara a la determinación de la conducta, de los estados físicos que configurarían la base de superveniencia de los estados mentales parece fuera de discusión, esto nos permitiría concluir sin mayores dificultades que los estados mentales gozan de la misma eficacia causal que los estados físicos en los que supervienen[331]. El problema, claro, es que el principio postulado es falso: hay infinitud de propiedades que pueden guardar un vínculo de superveniencia con una propiedad a la que atribuyamos eficacia causal y a las que –arguyen Jackson y Pettit (1990: 202)– en absoluto sería sensato adjudicársela también. En efecto, si tenemos una balanza diseñada para que un determinado circuito se cierre cuando el peso depositado en un platillo duplica al del otro, sobre el hecho de que así ocurra cuando en un platillo descansan tres gramos y en el otro seis no tiene vigor como causa que el peso en gramos en el primero equivalga a un número primo, que el segundo sea uno menos que siete, o que ambos sean divisibles por tres o por "[…] the weight of the Prime Minister […]", etc. –propiedades que, sin embargo, supervienen en las propiedades causalmente relevantes.

Pero lo que para Jackson y Pettit no parece suponer más que la expeditiva impugnación de un intento más bien desmañado de dotar de eficacia causal a lo mental esconde –cabe argumentar– frutos más valiosos. Bien podría decirse, ciertamente, que si de lo que se trata es de pergeñar una descripción de la cadena singular de causas y efectos que en una ocasión determinada ha hecho que se cierre el circuito de la balanza imaginada por Jackson y Pettit, nada –como no sea un mero canon de parsimonia– nos impide descartar como elementos de tal descripción cualesquiera propiedades que identifiquen unívocamente los pesos de cada platillo –como pongamos por caso, que en uno hubiera un peso menor de siete gramos en una unidad y en el otro un peso equivalente al menor divisor superior a uno del peso del Primer Ministro. Sólo cuando aspiramos a encerrar a ese evento causal singular en una generalización legaliforme capaz de respaldar enunciados contrafácticos –es decir, cuando queremos saber qué ocurrirá si los pesos en los platillos son, por ejemplo, ocho y cuatro gramos– las propiedades que elijamos para tipificar los eventos o estados involucrados –los pesos depositados en cada platillo– se tornan decisivas. Dicho de otro modo: al identificar los eventos o estados que intervienen en una relación causal mediante determinadas propiedades los tipificamos junto con otros que también las poseen –y no, se entiende, junto con otros con los que de hecho comparten sin duda algunas otras propiedades–; mientras nos mantengamos en el coto de la descripción de esa instancia particular de causalidad, el elemento de intensionalidad que esa tipificación introduce permanece inofensivo, pero tan pronto

[331] Estaríamos, si se quiere, ante una suerte de lectura inversa –o más bien, literal– del principio de herencia causal de Kim (1993a: 355, *supra*), generalizada para que concierna a propiedades, o las clases que conforman, más que eventos o estados particulares: que los poderes causales de un estado mental sean *idénticos* a los del estado físico en el que superviene, o en el que se instancia no significa que el estado mental carezca de poderes causales, sino más bien que posee exactamente los mismos que el estado físico –no posee poderes causales *propios*, podría decirse, pero tampoco entonces los posee el estado físico en cuestión.

como penetramos en la esfera de la generalización la intensionalidad inoculada por la propiedad elegida para caracterizar los eventos o estados involucrados en la relación causal comienza a ponerse de manifiesto. Esto, en cualquier caso, es una tesis que ya hemos tanteado reiteradamente (*cf*. Pylyshyn 1984: 4, Block y Fodor 1972b, *supra*): es en la intensionalidad de la explicación, por contraste con la extensionalidad de la descripción, donde puede buscarse el hábitat natural de una psicología autónoma, el nicho conceptual que podrían ocupar propiedades psicológicas explicativamente relevantes y, acaso, causalmente eficaces.

Donde Jackson y Pettit (1990) intentan que arraigue la relevancia explicativa del vocabulario psicólogico, en cambio, es en un análisis de la noción de propiedad funcional que lo asemeja a la de disposición –"[…] we can think of functional properties as more complex and general case of dispositional properties […]", acompañado de una reivindicación de la relevancia causal de las disposiciones que podamos mantener en pie por mucho que "[…] a full account of how […] events come about can be given in terms of the dispositions's categorical bases rather than the dispositions themselves" (Jackson y Pettit 1990: 203)[332]. En particular, la propuesta de Jackson y Pettit pasa por intentar apresar la relevancia causal de las propiedades disposicionales en las redes de la invarianza de sus efectos ante variaciones de su instanciación: si podemos decir –por abundar en el ejemplo que ellos mismos parecen preferir– que la conductividad eléctrica de un metal es relevante de cara a la explicación de un determinado desenlace –como que una persona sufra una descarga– es porque, primero, lo es la base categórica actual de la conductividad –digamos, la presencia de nubes de electrones libres– y, segundo, cualquier otra base categórica que hubiera preservado el patrón de relaciones causales que define la conductividad eléctrica habría provocado idéntico desenlace. Así, apuntan Jackson y Pettit (1990: 205), "[…w]e move from the non-contentious causal relevance of the categorical basis to the causal relevance of the disposition"[333].

---

[332] La cuestión de la eficacia causal de las disposiciones, por oposición a la de su base categórica, ha sido ya abordada en relación con la controversia entre Ryle (1949) –o, más nítidamente, Price (1953)– y Geach (1957). Es significativo, sin duda, que ciertos aspectos de la defensa del carácter causal de las disposiciones, en la que el conductismo lógico hubo de atrincherarse entonces ante los embates de un funcionalismo que reclamaba el papel causal de los estados internos, reaparezcan luego bajo la forma de una defensa, precisamente, del papel causal de esos estados internos, entendidos como estados funcionales, ante la pujanza, esta vez, de los apelación a propiedades físicas de los estados en los que se encarnan.

[333] No es del todo cierto, como hemos visto, que la idea de que la relevancia causal compete a la base categórica de la disposición no revista ninguna controversia. Contra ella, por ejemplo, pesa el ensayo de refutación que Squires (1968) presentara de la tesis de Armstrong (1968: 86, *supra*) según la cual son ciertas propiedades físicas de los cuerpos frágiles –la base categórica de la fragilidad– lo que causa efectivamente su rotura en ciertas circunstancias, y no, propiamente, la fragilidad: comoquiera que bien puede darse el caso de que esas circunstancias nunca se den, y no querríamos por ello concluir que el objeto en cuestión no es frágil, "[…] Armstrong can only mean that the [categorical] basis would cause the object to behave in the relevant ways in appropriate circumstances"; pero entonces –claro está– se está atribuyendo a la base categórica una propiedad disposicional "[…] suspiciously similar to that which was originally attributed to the object itself" (Squires 1968: 45). No sería

La dificultad que Jackson y Pettit tratan de resolver radica en que la misma base categórica que subyace a la conductividad eléctrica lo hace también –aseguran ellos mismos: *cf*. Jackson y Pettit (1990: 204)– a otras muchas propiedades disposicionales, como la conductividad térmica o la opacidad, que en cambio bien pueden no ser relevantes en la explicación de un determinado evento. El criterio que nos permite excluirlas de la explicación es, naturalmente, que de darse idéntica disposición con otra base categórica distinta –que no acarrease también conductividad eléctrica– no se hubiera producido el desenlace en cuestión: así, si el material hubiera sido madera habríamos tenido opacidad pero no conductividad y tampoco, desde luego, descarga eléctrica. El curso de la argumentación, por tanto, parece dejar claro que, de nuevo, es en el ámbito de la generalización y lo contrafáctico, no en el de la descripción de un evento particular, donde las propiedades funcionales, o disposicionales, adquieren su valor explicativo:

> The explanatory interest of an explanation in terms of a dispositional property is now clear. We are often interested not merely in how something in fact came about but also in how it would have come about. (Jackson y Pettit 1990: 205)

No menos claro resulta, sin embargo, que la ruta señalada por Jackson y Pettit sólo puede conducir a una noción de relevancia explicativa despojada de todo sustrato de eficacia causal autónoma –como ya se ha visto en relación con Braddon-Mitchell y Jackson (1996)[334]. Lo que la apelación a propiedades funcionales nos procura, como lo hace la apelación a disposiciones, es un grado superior de abstracción –un modo de incrementar la potencia de una explicación "[…] by, in a sense, saying less" (Jackson y Pettit 1990: 205). La historia es conocida:

> A certain piece of behaviour will have a certain property, say that of being in the direction of a certain cup of coffee, as a result of the concatenation of very many neurophysiological states which will have given rise to that piece of behaviour by virtue

---

aventurado afirmar que lo que la amenaza de *regressus* esconde es, una vez más, la dificultad de construir una noción robusta de concatenación singular de causa y efecto al margen de una malla nomológica erizada de contrafácticos –o, cuando menos, una noción de explicación en términos de tales concatenaciones singulares.

[334] La apariencia, en el ejemplo de la conductividad y la opacidad, de que la disposición pudiera albergar una eficacia causal propia, diferenciada de la que atañe a su base categórica, es ilusoria, pues parte de la afirmación equívoca de que la base categórica de ambas propiedades disposicionales es idéntica cuando de hecho –contrastamos después– la opacidad puede reposar en propiedades físicas que no conllevan conductividad –como ocurre en el caso de la madera. Así, aunque la conductividad eléctrica pueda ser relevante en la explicación de una descarga –no siéndolo, en cambio, la opacidad–, en el análisis de Jackson y Pettit la eficacia causal es privativa de las propiedades físicas de la nube de electrones libres que se forma en los enlaces metálicos.

Esa laxitud en la identificación de la base categórica de las disposiciones, o de la instanciación física de las propiedades funcionales, parece consustancial, por cierto, a la defensa "impenitente" (Jackson y Pettit 1990: 199) de que cada tipo de estado mental, aun cuando sea de hecho definido en términos funcionales, se identifica con un tipo de estado neurofisiológico, por mucho que éste pueda no ser el mismo en distintas especies, sujetos, o instantes.

> of their natures, that is, by virtue of the neurophysiological properties they instantiate. But, of course, there will be other ways that behaviour with the property of being towards the coffee could have been caused, other neurophysiological ways, or even, other non-neurophysiological ways if we allow ourselves Martian speculations. Is there anything interesting that we can say about resemblances between these various actual and possible ways of getting behaviour towards the coffee? The answer is that it may be that many of these ways, including the actual way, are united by the functional properties they realize, and in particular by the functional properties definitive of contents that they realize. In that case, an explanation in terms of content-bearing states will apply and its explanatory interest will lie in the fact that it tells us about what would happen in addition to what did happen. That is how the content properties may be causally relevant. (Jackson y Pettit 1990: 206)

Pero el deseo de –supongamos– tomar un café no hace nada para provocar los movimientos del brazo: todo lo hace esa intrincada concatenación nerviosa, en virtud de las propiedades neurofisiológicas de los estados que la integran. Según lo plantean los propios Jackson y Pettit (1990: 207), anticipándose a la objeción de que su táctica no elude la espada de Damocles del epifenomenismo:

> But all that that shows is […] the explanatory value of content ascriptions. It does not show that content properties conceived functionally do the driving of behaviour. The fact remains that that is done by neurophysiological (or least relatively intrinsic structural or syntactic) properties; yet surely the commonsense intuition that cries out for vindication is that content drives behaviour.

La respuesta que Jackson y Pettit ensayan para esa objeción se adentra de lleno en la metafísica de la causalidad. Si asumimos una concepción mecánica de la relación entre causa y efecto en detrimento de una escuetamente nomológica, humeana –tal como hemos visto que ocurre de forma más o menos velada en Kim, Bickle, Toribio o Fodor–, si, es decir, "[…] we think of causation as a matter of production or efficacy which does not reduce sooner or later to nothing more than nomological sequence" y entendemos que "[…] according to this view, a sequence is nomological because of underlying causal productivities, not conversely" (Jackson y Pettit 1990: 208), la preocupación por el estatus causal de las propiedades psicológicas se desvanecerá por si misma[335]. La razón es sencilla: esa concepción de la causalidad, al fin y al cabo,

[335] Al paso dejan apuntado además Jackson y Pettit (1990: 208) que tal idea de la causalidad resulta obligada si hemos de admitir la existencia de causas rigurosamente singulares, "[…] in the sense of one event causing another which does not fall under a law, either deterministic or indeterministic", como han defendido Anscombe (1975) o Tooley (1990). El debate sobre la naturaleza de la causalidad al que nos ha abocado el esfuerzo por entender ciertas peculiaridades de la explicación psicológica abre el camino, así, a una reflexión acerca de la propia idea de orden –o de cosmos, si se quiere acentuar las reverberaciones pitagóricas de la cuestión: la pregunta es, entonces, si un mundo en el que existen causas estrictamente singulares, no humeanas, es un mundo constitutivamente desordenado y, en consecuencia, acaso constitutivamente incomprensible.

Una idea mecánica de causalidad como la que enarbolan Jackson y Pettit, sin embargo, puede construirse sin necesidad de aceptar que existan eventos causales estrictamente singulares. Ni que

"[…] enjoins us to restrict relations of causal efficacy to certain properties in fundamental Physics […] and to see all the causal relevancies 'higher up' as, strictly speaking, non-efficacious". Así que, después de todo, fisiología y psicología quedan empatadas, habida cuenta de que "[…] neurophysiological properties are not causally efficacious in the special sense any more than […] content properties" (Jackson y Pettit 1990: 209). Mal de muchos, ya se sabe:

> […] the functionalist account of content does not downgrade its causal role, rather it leaves it in the excellent company of everything except for certain members of that most exclusive of clubs, the properties of fundamental physics. (Jackson y Pettit 1990: 210)[336]

Tal conclusión, por cierto, no parece del todo ingrata para Fodor (1989), quien, tras advertir que el argumento que supuestamente nos aboca a una interpretación epifenomenista del funcionalismo –a la vista de que se aplica por igual a cualquiera de las propiedades que puedan postular las ciencias especiales– "[…] *has nothing to do with intentionality as such*" (Fodor 1990: 61), apunta que:

> […] there are likely to be parallel arguments that *all properties are causally inert except those expressed by the vocabulary of physics. In which case, why should anybody care* whether psychological properties are epiphenomenal? All that anybody could reasonably want for psychology is that its constructs should enjoy whatever sort of explanatory/causal role is

---

decir tiene que Fodor rechaza de plano la idea de causa singular, al menos en el plano epistemológico: "I assumme that singular causal statements need to be covered by causal laws" (Fodor 1989: 64).

[336] Parece quedar en el aire, en el razonamiento de Jackson y Pettit, la sospecha de que el círculo de la eficacia causal podría acaber siendo tan exclusivo que no logremos dar con sus legítimos integrantes:

> […] we do not need to believe in any fundamental efficacies over and above those between properties at the micro-level in order to explain the regularities, actual and counterfactual, all the way up, because supervenience tells us that they are fixed by how things are at the bottom (*if* there is a bottom). (Jackson y Pettit 1990: 209)

La misma duda aflora en Block (2003, *supra*): que lo que hemos dado en llamar "poderes causales" termine enteramente drenado, sin residuos, en el hallazgo de propiedades físicas subyacentes a cualesquiera pudiéramos tomar por elementales. La idea, en cambio, de que la causalidad es *stricto sensu* una relación entre fuerzas –y desde luego "[…] a robust ingredient within the world itself"– es defendida por Bigelow y Pargetter (1990), a quienes Jackson y Pettit se remiten a pie de página.

Parecen traslucir también inquietudes parejas a estas cuando Fodor describe como "[…] a bold assumption" la idea de que todos los mecanismos que implementan las leyes de las ciencias especiales son de naturaleza física, para añadir a renglón seguido la confesión de que desconoce "[…] what this bold assumption means […], and […] exactly why it seems to me to be a reasonable bold assumption to make", toda vez que ignora también "[…] what it *is* for a mechanism to be physical […]" (Fodor 1989: 76). De ahí que, a fin de cuentas, Fodor (1989: 77) asegure sentirse más proclive –de verse entre la espada y la pared, claro– a renunciar al fisicalismo que a renunciar al realismo sobre el papel de los estados mentales como causas de la conducta –lo que, como ya sabemos, identifica entre bromas y veras ni más ni menos que con el fin del mundo (Fodor 1989: 77, Fodor 1990: 156).

> proper to the constructs of the special sciences. If beliefs and desires are as well off ontologically as mountains, wings, spiral nebulas, trees, gears, levers and the like, then surely they're as well off as anyone could need them to be. (Fodor 1989: 63)[337]

Claro que, con todo –discrepa Fodor a renglón seguido de esa provisional conformidad– "[…] something must have gone wrong with arguments that show that all these properties are epiphenomenal": de lo contrario, sencillamente no podrían existir ciencias especiales[338]. Lo que procede, pues, es tratar de construir una noción de eficacia causal –"responsabilidad causal", dice Fodor (1989)– que no expulse de su seno a las propiedades de las ciencias especiales. O, como Fodor (1989: 63) prefiere describir su proyecto, "[…] a tonic for epiphobics".

Cabría entonces, quizás, intentar zafarse del dictamen de Jackson y Pettit aprovechando más esa honda distensión que –decíamos– injiere la tesis de realizabilidad múltiple en los vínculos entre las propiedades psicológicas a las que apelamos en la explicación de tal o cual conducta y su sustrato físico, y tratando de ensanchar ese margen de maniobra mediante una elucidación del papel de las cláusulas *cæteris paribus* en las generalizaciones que integrarían el cuerpo teórico de la psicología. Ésa es la estrategia del propio Fodor (1989, 1991), quien responde así a las objeciones que Schiffer (1991) había presentado contra la idea de que puedan existir leyes que expresen regularidades *cæteris paribus* –básicamente, porque el contenido de la cláusula precautoria resulta ser imposible de especificar, con lo que la noción de una ley *cæteris paribus* vendría al fin y al cabo a ser, tal como lo plantea Fodor (1991: 21), un oxímoron–, o de que sean necesarias tales leyes para preservar la relevancia explicativa de la apelación a estados internos en las teorías psicológicas[339].

---

[337] Cabe argumentar, no obstante, que a una defensa de la autonomía de la psicología le bastaría con recibir idénticos parabienes que cualquier otra ciencia especial si su objetivo fuera recusar la reducción de su estructura teórica a la de la física; como el debate fundamental, en cambio, no es de hecho si cabe reducir la psicología a física sino a fisiología, acaso sean precisos criterios sólidos para distinguir la explicación psicológica no ya, genéricamente, de las leyes fundamentales de la física, sino de otras leyes especiales, como las que se pudieran establecerse por medio del vocabulario teórico de la neurofisiología. El carácter semántico, intencional de las propiedades a las que se apela en la explicación psicológica parece un pretendiente nato para desempeñar ese oficio, siempre y cuando logremos por fin dotarlo de la autonomía causal, *ergo* explicativa, en torno a la que orbitan estas reflexiones.

[338] Acaso valga la pena señalar que existe un nítido paralelismo entre la línea argumentativa que esgrime aquí Fodor y la que en pugna con el eliminacionismo articulara Searle (1992: 47, *supra*) cuando arrancaba de éste la poco decorosa conclusión de que en realidad no existen cosas tales como "[…] golf clubs, tennis rackets, Chevrolet station wagons, and split-level ranch houses".

[339] El estatus epistemológico de las cláusulas *cæteris paribus* es un asunto candente en el ámbito de la filosofía de la ciencia –particularmente, desde que ya en 1890 los *Principios de Economía* de Alfred Marshall las hicieran proliferar aquí y allá, en la filosofía de la economía– cuya importancia para la filosofía de la psicología se ha visto muy realzada a partir de este intercambio entre Schiffer (1991) y Fodor (1991). Abordarlo con la profundidad que merecería queda fuera de los propósitos del presente estudio.

En efecto, si asumimos que los estados mentales son estados funcionales cuya implementación física puede variar en distintos sistemas, o en el mismo sistema a través del tiempo, y que dichos estados mentales intervienen en generalizaciones acerca de la conducta del sujeto que los alberga, generalizaciones que incorporan cláusulas *cæteris paribus*, podemos, con Fodor, desgranar esquemáticamente el contenido de dichas cláusulas de la siguiente forma: la conjunción de una determinada encarnación física del estado funcional en cuestión y unas ciertas condiciones adicionales son suficientes para provocar tal o cual conducta, pero ni uno ni otro término de la conjunción es suficiente para ello por sí solo; a esas condiciones adicionales es a las que alude tácitamente la cláusula *cæteris paribus*. Provistos de este croquis podríamos entonces distinguir tres clases de generalizaciones en función del tipo de excepciones que admitan –entendiendo de entrada como excepción un suceso que cae "[…] under the antecedent of the law and is not the cause of an event that falls under the consequent" (Fodor 1991: 24). Tendríamos en primer lugar las leyes estrictas –como serían, se entiende, las leyes fundamentales de la física–, que no admiten excepciones: toda excepción constituiría evidencia contraria a la ley[340]. En segundo lugar habría que consignar las leyes *cæteris paribus* –como, se entiende, las de la psicología. Estas leyes admiten *meras excepciones*, en las que se da un determinado estado físico que constituye una encarnación del estado funcional citado en la ley, pero las condiciones adicionales no se dan, así como –piensa Fodor– *excepciones absolutas*, en las que tales condiciones no pueden siquiera llegar a fijarse dado que sencillamente no existe un conjunto definido de circunstancias en conjunción con las cuales el estado físico en cuestión genere de cierto la conducta. Lo decisivo en este caso es a qué clase de conceptos apela la ley en cuestión: si se trata de conceptos cuya definición viene dada en términos funcionales, entonces las excepciones absolutas son posibles, porque puede darse el caso de que para algunos de los estados físicos en los que se encarna el estado funcional que tal concepto identifica no existan condiciones adicionales que aseguren el cumplimiento de la ley, pero sí para otros de esos estados físicos; en cambio, si tuviéramos una ley *cæteris paribus* que no estuviera formulada en términos de conceptos funcionales, sino de conceptos que fijaran directamente el mecanismo subyacente –en términos de "esencias ocultas", dice Fodor (1991: 30)–, las excepciones absolutas no serían viables –o, mejor dicho, no serían excepciones, sino evidencia contra la verdad de la ley[341].

[340] Distinta cuestión, naturalmente, es que dichas leyes puedan venir formuladas en términos probabilísticos (*cf*. Schiffer 1991: 1, Fodor 1991: 21, 30), lo cual en nada parece, por lo demás, que afecte al argumento.

[341] Más pormenorizadamente:

> The point here is that, whereas all you need for a mere exception is a realizer that is tokened without its completer, you only get absolute exceptions when some realizers have completers and others don't. (A state *none* of whose realizers have completers for a given law is simply a state that doesn't fall under the law.) So, by definition, you only get absolute exceptions to laws about multiply realized kinds. But, again by definition, a kind with a hidden essence has *only one* type of

Por último, si hemos de considerarlas generalizaciones, estarían también las generalizaciones vacías –o, cabría decir, pseudogeneralizaciones–, en las que se dan excepciones absolutas que lo son no sólo respecto de la propia generalización, sino respecto de cualquier otra generalización en la que aparezca como antecedente el estado funcional en cuestión –digamos, así pues, *excepciones radicales*, o, con Fodor (1991: 27), excepciones absolutas generalizadas, "across the board".

La clave del argumento de Fodor reposa en un ingenioso giro que le permite vetar la conclusión de que las leyes psicológicas, en las que aparecen como antecedentes estados mentales definidos funcionalmente, puedan ser en realidad pseudogeneralizaciones y no leyes *cæteris paribus*. Para que una ley psicológica resultara ser una generalización vacía, habría de darse el caso, según hemos visto, de que un determinado estado físico en el que se encarnara el estado funcional que figura en el enunciado de la ley constituyera una excepción radical –es decir, una excepción absoluta no sólo a dicha ley, sino al conjunto íntegro de leyes psicológicas en las que el estado funcional en cuestión figurase. Pero aquí –arguye Fodor (1991: 28)– "[…] hemos tocado fondo": comoquiera que en una ley psicológica los estados internos involucrados vienen definidos en términos funcionales, si tal cosa ocurriera sencillamente no tendríamos ninguna razón para admitir al estado físico que supuestamente convertiría la ley en una pseudogeneralización como encarnación del estado funcional en cuestión. Dicho de otro modo: si cada vez que dicho estado físico se da en un sujeto nos topamos con que no se produce ninguna de las conductas que las leyes psicológicas vinculan a un determinado estado mental, definido funcionalmente, y ni siquiera es posible fijar unas condiciones adicionales bajo las cuales el estado físico en cuestión provocaría alguna de dichas conductas, entonces no hay desde luego ningún sustento para afirmar que los sujetos que se encuentran en dicho estado físico alberguen *eo ipso* el estado mental que nos ocupa[342]. Al contrario:

> […] you can imagine a state for which the functional definition of the realizer relation is compatible with absolute exceptions to *any given* law in a network, so long as *most* of the laws are satisfied. But to find a state that is an absolute exception to all the laws that subsume a certain attitude *just is* to find a state that doesn't realize that attitude. You couldn't ask for better evidence. (Fodor 1991: 29)

Precisamente, pues, en la medida en que las leyes *cæteris paribus* que forman el cuerpo teórico de la psicología se formulan apelando a estados funcionales –es decir, estados cuya relación con los estados físicos que los encarnan queda definida por la

---

> realizer; so, trivially, either all of its realizers have nomologically possible completers for a given law or none of them do. (Fodor 1991: 30-31)

[342] *Cf*. Lewis (1980, *supra*) sobre la posibilidad de concebir un estado de dolor enteramente desgajado de la malla de causas y efectos en la que suele entretejerse. Ya entonces se adelantaba, en efecto, que Fodor debe rechazar que lo imaginado por Lewis sea verdaderamente concebible –como hacía Putnam (1967a: 229, *supra*)–, o bien conceder que el dolor no sea un estado funcional.

urdimbre nomológica en la que se imbrican las diversas instancias de los propios estados funcionales (*cf*. Fodor 1991: 29)–, la posibilidad de que esas leyes no resulten ser más que pseudogeneralizaciones se desvanece ante nuestros ojos. Al mismo tiempo, aun si asumimos que lo que hace verdaderas tales leyes, como las de cualquier ciencia especial, no es sino el trabajo de mecanismos físicos –"[…] non-basic laws want implementing mechanisms; basic laws don't", asegura Fodor (1989: 76; *cf*. Fodor 1994b: 294)[343], y, de hecho, el laborioso avance de las ciencias especiales suele consistir a su juicio precisamente en "[…] an interplay between the discovery of higher level laws and the discovery of lower level implementations" (Fodor 1991: 20)[344]–, la relevancia explicativa de las leyes psicológicas, y con ella la de los conceptos referidos a estados mentales que aducen, quedaría asegurada por el hecho de que no es posible formular una ley que sin emplear tales conceptos capture la misma generalización, dado que el mecanismo físico que encarna cada instancia en que la generalización se cumple puede ser indefinidamente diferente. En otras palabras, lo que las cláusulas *cæteris paribus* nos proporcionarían sería, seguramente entre otras cosas, la posibilidad de cuantificar sobre esos mecanismos subyacentes:

> Nonbasic laws *rely on* mediating mechanisms which they do not, however, *articulate* (sometimes because the mechanisms aren't known; sometimes becauses As can cause Bs in many different ways, so that the same law has a variety of implementations). Ceteris

[343] Asunción que, como hemos visto, también Toribio (1991: 5, *supra*), da por buena. En esa diferencia entre leyes básicas y especiales cifra Fodor (1989: 76) el carácter *básico* de aquellas. Más pormenorizadamente:

> […] a metaphysically interesting difference between basic and nonbasic laws is that, in the case of the latter but not of the former, there always has to be a *mechanism in virtue of which* the satisfaction of its antecedent brings about the satisfaction of its consequent. (Fodor 1989: 66).

[344] Aun cuando no son estrictamente equivalentes, cabe expresar la convicción de Fodor de que a toda ley psicológica subyace un mecanismo físico por medio de la idea de que "[…] special science laws […] are characteristically 'heteronomic'" (Fodor 1989: 78) –es decir, no podrían convertirse en leyes estrictas sin recurrir al vocabulario de ciencias más básicas.

La distinción entre generalizaciones heteronómicas y homonómicas, tal como se emplea hoy en filosofía de la psicología, tiene su origen en Davidson (1970: 219):

> [...o]n the one hand, there are generalizations whose positive instances give us reason to believe the generalization itself could be improved upon by adding further provisos and conditions stated in the same general vocabulary as the original generalization. Such a generalization points to the form and vocabulary of the finished law: we may say that it is a *homonomic* generalization. On the other hand, there are generalizations which when instantiated may give us reason to believe there is a precise law at work, but one that can be stated only by shifting to a different vocabulary. We may call such generalizations *heteronomic*.

Como también mantendrá Fodor, Davidson considera que las leyes psicológicas (así como las de orden psicofísico: las que enlazan propiedades mentales con propiedades físicas) son siempre heteronómicas; a su juicio, esto es tanto como decir que no pueden ser leyes estrictas. Que la distinción entre generalizaciones heteronómicas y homonómicas sirva para respaldar el monismo anómalo defendido por Davidson es cuestión muy discutida: *cf*. a modo de síntesis Yalowitz 2005.

> paribus clauses can have the effect of existentially quantifying over these mechanisms, so that "As cause Bs ceteris paribus" can mean something like "There exists an intervening mechanism such that As cause Bs when it's intact." (Fodor 1989: 76)

Así pues, incluso si el hecho de que un determinado estado psicológico que hemos definido en términos funcionales cause una cierta conducta puede quedar *en cierto sentido* explicado por el hecho de que el estado físico en que se encarna tenga precisamente esos efectos de acuerdo con una ley física –es decir, que la relación entre el estado psicológico y la conducta estaría sobredeterminada (*cf.* Fodor 1991: 30)– , las leyes psicológicas en las que figura dicho estado funcional pueden resultar irreemplazables –diría Fodor– a la hora de dar cuenta de la generalización según la cual diversos estados físicos comparten los efectos en cuestión –es decir, a la hora de explicar, *en otro sentido*, que instancias de ese estado psicológico funcional causen la conducta: que lo hagan con independencia del estado físico en que se encarnen. La defensa de la autonomía de la explicación psicológica largamente enarbolada por Fodor tendría su piedra angular, entonces, en la distinción entre dos sentidos de la idea de explicación o, si se prefiere, en la diferenciación entre dos funciones que las leyes desempeñan en nuestros esquemas explicativos:

> […] laws do more than cover singular causal relations. *They also function to express generalizations* which may not be able to be captured in any except their proprietary vocabulary. It's important to see that there may be only one way to express a certain generalization, even though there is more than one way to cover a certain causal relation. Or, what comes to much the same thing, the function of laws in explaining and predicting events that fall under them need not be its *whole* function. A law may be essential for capturing generalizations even where overdetermination would, in principle, make it superfluous for causal explanation. When this is the case, we put up with the overdetermination because we want the generalization. (Fodor 1991: 30)

En definitiva –piensa Fodor– no hoy motivos para alarmarse o para proclamar un nuevo esplendor del reduccionismo, pues aunque "[…] some philosophers have thought they saw a conflict between the idea that psychological explanations are typically intentional and the idea that they are typically computational", hay buenas razones para pensar que "[…t]hat, however, is a mistake" (Fodor 1991: 19-20).

Es, así pues, la diferencia entre *explicar* en tanto que *dar cuenta de una relación causal singular* y *explicar* en tanto que *apresar una generalización sobre relaciones causales* lo que de acuerdo con la propuesta de Fodor dota de autonomía a las leyes que involucran estados psicológicos[345], ya que mientras *dar cuenta de una relación causal singular* entre un estado psicológico y una conducta puede hacerse sin alusión al estado psicológico como tal –recurriendo sólo a las leyes físicas que especifican el

[345] Se trata, entonces, de lo que Pylyshyn (1984: 4, *supra*) caracterizaría más bien como la diferencia entre *describir* y *explicar* –entre la extensionalidad de aquella labor y la intensionalidad de ésta–, a la que subyace, como aquí, la diferencia entre instancias singulares de eventos o de relaciones causales y clases de eventos o de relaciones causales, y que el propio Pylyshyn –recuérdese– hace arraigar en ciertas observaciones de Block y Fodor (1972b).

mecanismo causal subyacente a esa instanciación singular de la generalización expresada en la ley–, *apresar una generalización sobre relaciones causales* entre estados psicológicos y conductas sólo es posible, toda vez que los estado funcionales pueden venir encarnados de modos indefinidamente diversos, si contamos con el arsenal teórico de la psicología.

La ambigüedad en el uso de la noción de explicación a la que apuntaba el argumento de Toribio (1991, *supra*) parece, de esta forma, quedar elucidada. Recordemos, con todo, que el problema surgía cuando topábamos con las conclusiones contrapuestas de que, por un lado, la descripción física no es relevante para la explicación de los poderes causales de las propiedades semánticas de los estados mentales –o digamos, más modestamente, no la agota– en la medida en que dichos estados mentales vienen definidos en virtud de propiedades funcionales y, por otro, de que *los poderes causales de las clases de eventos* definidas en virtud de propiedades macrofísicas –esto es, por ejemplo, de propiedades funcionales– quedan explicados por los poderes causales de los eventos microfísicos particulares en los que cada elemento de dicha clase superviene (Toribio 1991: 5, supra). El razonamiento de Fodor parece invitarnos a considerar la siguiente resolución de la contradicción que observábamos en Toribio: decimos que la descripción física no agota la explicación de los poderes causales de las propiedades semánticas de los estados mentales en tanto que no permite *apresar todas las generalizaciones* sobre relaciones causales en las que dichos estados pueden verse trabados por razón de sus propiedades semánticas, y decimos que los poderes causales de las clases de eventos definidas en virtud de propiedades de propiedades funcionales quedan explicados por los poderes causales de los eventos físicos particulares en los que cada elemento de dicha clase superviene en tanto que *toda relación causal singular que pueda formar parte de la definición funcional de dicha clase de eventos queda cubierta* por la explicación física.

Pero es importante advertir que el sentido en que se afirma que el vocabulario de la física proporciona una explicación completa no tiene que ver con eventos causales singulares, sino precisamente con *los poderes causales de las clases de eventos* definidas en virtud de propiedades funcionales: es el hecho de que una clase funcional de eventos –por ejemplo, de eventos de naturaleza mental– vea sus poderes causales explicados por los de ciertos eventos físicos particulares –con los que, recuérdese, se pretende que guarde superveniencia aunque estos sean diferentes en diferentes ocasiones– lo que Toribio (1991, *supra*) toma como garantía de que cabe atribuir eficacia causal a la clase funcional. Naturalmente, cabe elevar una propuesta de corrección: no son en realidad los poderes causales de las clases de estados mentales –de las clases de eventos definidas en función de propiedades macrofísicas, funcionales– los que quedan explicados por los poderes causales de las instancias de estados físicos en los que cada elemento de la clase superviene, sino los poderes causales de los estados mentales particulares que forman la clase funcional. La diferencia es relevante precisamente porque dichos estados mentales particulares pueden encarnarse en distintos estados físicos, con poderes causales indefinidamente

distintos, en distintas ocasiones, por lo que no es cierto que los poderes causales de la clase funcional queden explicados por los de los estados físicos en los que se encarnan sus elementos –como no sea que queramos admitir como explicación una disyunción indefinidamente prolija y dispar.

Se diría, de hecho, que sin trazar la distinción entre los efectos que podamos atribuir a un estado mental singular y los poderes causales que quepa adjudicarle a la clase funcional a la que pertenece, la posición de Toribio parece revertir sin remedio hacia la de Jackson y Pettit (1990), donde, como sabemos, la eficacia causal es privativa de los eventos físicos singulares. Se trata, en suma, de la concepción de la causalidad cuyas resonancias epifenomenistas respecto de lo mental Toribio trata de esquivar sin cuestionarla, y que condensa así:

> Only individual events described from a micro-physical point of view can be causes. Now, causal stories at a higher level don't unfold in terms of singular events, but generalise over conjunctions and disjunctions of lower level events. Therefore, the causal laws that belong to these higher levels can only be true in virtue of the causal efficacy of the micro-physical events which the terms of the macro-physical or intentional vocabulary that appears in those laws refer to in the end. (Toribio 1991: 6)

Al distinguir entre la labor de rendir cuentas de una concatenación singular de causa y efecto, para la cual bastaría el recuento de las propiedades físicas de éste y aquélla, y la de apresar determinadas generalizaciones en las que se ven envueltas clases de eventos cuyas encarnaciones físicas pueden ser sumamente heterogéneas, para la cual nos sería preciso el vocabulario funcional que nos permite delimitar la clase en cuestión, se trataría al fin y al cabo de restaurar un margen de eficacia causal autónomo para las propiedades a las que alude nuestro vocabulario funcional. Es a esto a lo que parece apuntar el propio Fodor cuando de forma más taxativa, con tono un tanto airado, intenta dejar claro su rechazo a la idea de que la semántica sea prescindible en la teorización psicológica:

> THE CLAIM THAT MENTAL PROCESSES ARE SYNTACTIC DOES NOT ENTAIL THE CLAIM THAT THE LAWS OF PSYCHOLOGY ARE SYNTACTIC. On the contrary THE LAWS OF PSYCHOLOGY ARE INTENTIONAL THROUGH AND THROUGH. […] What's syntactic is not the laws of psychology but the mechanisms by which the laws of psychology are implemented. Cf: The mechanisms of geological processes are –as it might be– chemical and molecular; it does not follow that chemical or molecular properties are projected by geological laws (on the contrary, it's geological properties that are projected by geological laws); and it does not follow that geological properties are geologically inert […]. (Fodor 1989: 67, mayúsculas en el original)

Todo bien, entonces. Pero no deja de ser interesante observar que cierta apariencia de velada contradicción persiste, no obstante, también en las conclusiones del propio Fodor (1991), y por motivos muy parecidos. Justo antes de asegurar que las leyes científicas, además de cubrir relaciones causales singulares, "[…] *also function to express generalizations* which may not be able to be captured in any except their

proprietary vocabulary" (Fodor 1991: 30, *supra*), el propio Fodor concede que las explicaciones de las ciencias especiales posiblemente estén sobredeterminadas, ya que:

> If an event falls under the antecedent of a special science law, it must have some property in virtue of which it also falls under the antecedent of a basic science law and such that the consequent of the special law is satisfied whenever the antecedent of the basic law is satisfied. (Fodor 1991: 30)

Pero si esto es cierto, bien parece que toda generalización que pudiera apresar la ley especial en cuestión caería también en las redes de la ley básica que, así las cosas, confiere su vigor a la ley especial. No podríamos entonces, o no deberíamos –como sugiere Fodor (1991: 30, *supra*) que en la práctica hacemos– asumir la sobredeterminación a cambio de la generalización que se nos procura, pues no se nos procura ninguna que no tuviéramos ya. Como si reparara en la inestabilidad de su posición, Fodor emprende en este punto un expeditivo trámite encaminado a ensanchar aun el espacio nomológico en el que las propiedades psicológicas se incardinan.

> All this is true in spades when we want to express not just an empirical generalization which may subsume a multiply realized kind, but also the membership of that generalization in a network of laws with partially overlapping antecedents; or when we wish to define a functional state relative to the whole population of laws that constitutes a network […]. There is, to come to cases, no reason to suppose that any projectible predicate of a nonpsychological science will collect the instantiations of a psychological state over the whole network of laws that control it. The sum and substance is the truism that we need psychological vocabulary to say what the realizers of a psychological state have in common. (Fodor 1991: 30)

No es ya, por consiguiente, el hecho de que eventos físicamente dispares puedan quedar amparados por una misma ley cuando se taxonomizan en virtud de propiedades psicológicas –funcionales– de forma tal que no haya ley física que exprese la misma regularidad, pues la existencia de leyes básicas que reflejan cualquier regularidad que pueda expresar una ley especial esto ha quedado concedida. La cuestión se traslada, más bien, a la tesis de que esos eventos físicamente dispares se entretejen en una urdimbre de leyes, y su presencia en unas u otras sólo podría quedar explicada aludiendo a las propiedades psicológicas que pese a su heterogeneidad física comparten. No es fácil ver, sin embargo, cómo podría la imbricación de un evento físico en un conjunto de leyes especiales no quedar explicada por su imbricación en un conjunto de leyes básicas toda vez que su vinculación a cada ley especial queda explicada por su vinculación a una ley básica. Lo mismo ocurre si examinamos las indicaciones del tónico contra la *epifobia* que Fodor (1989) nos receta, cuyo ejercicio crucial es poner todo el cuidado en evitar dos confusiones:

> It's a confusion to suppose that, if there's a law, then there needn't be an implementing mechanism; and it's a confusion to suppose that, if there's a mechanism that implements a law, then the properties that the law projects must be causally inert. (Fodor 1989: 68)

Ahora bien: si hay *un* mecanismo fisiológico –sintáctico, si se prefiere– que implementa cada ley psicológica –que aluda a propiedades semánticas de estados mentales–, entonces, nos guste o no, es difícil, nuevamente, encontrar motivos para adjudicar responsabilidad causal a los estados mentales aludidos o a sus propiedades semánticas, habida cuenta de que cualquier cosa que la ley en cuestión pueda explicar o predecir la explicará o predecirá también la descripción del funcionamiento del mecanismo fisiológico que necesariamente implementa la ley.

En suma, los esfuerzos de Fodor por afianzar la eficacia causal de las propiedades psicológicas, como los de Toribio, sólo llevan a buen puerto –cabe concluir– si media la renuncia a la idea de que los poderes causales que cabe atribuir a una propiedad de orden superior –o a una clase de estados o eventos definida por la instanciación de dicha propiedad– se agotan en los que quepa atribuir a las propiedades físicas de los eventos o estados que conformen los elementos de la clase que la propiedad de orden superior delimita[346]. Claudicar de esa idea, ahora bien, parece vedado por la convicción de que toda relación causal es en último término una relación entre eventos descritos en la más elemental de las teorías físicas, y que se despliegan –como ya se ha dicho– en el ámbito de los constituyentes últimos de la materia[347].

---

[346] Lo que está en el aire de nuevo, así pues, no es sino la interpretación del principio de herencia de poderes causales que Kim (1993a: 355, *supra*) considera definitorio de la relación de realización –si bien en la formulación de Kim, como hemos visto, el principio atañe sólo a los poderes causales de eventos o estados particulares, o, si se quiere, propiedades instanciadas en ellos, no a los de propiedades en tanto que criterios que definen clases o tipos de estados o eventos.

[347] Algunos trabajos posteriores de Toribio bien pueden leerse, no obstante, como una exploración de diversos itinerarios que pudieran conducir a la idea de explicación psicológica a liberarse de la servidumbre de la sintaxis a la que la unce la idea de que "[…] intentional states require identifiable inner vehicles […] " (Toribio 1998: 129) y de que son las propiedades intrínsecas de esos vehículos las que *constituyen* el estado mental en cuestión. Cuando lo que importa –dice Toribio– es el armazón de interacciones con el entorno, y en especial de prácticas sociales, que fija, en un sentido normativo, a un estado mental o una expresión lingüística con su significado, la noción de constitución deja de ser adecuada y debemos pensar más bien –en la estela del último Wittgenstein– en términos de institución. En el intento de Brandom (1994) de desplegar una concepción del significado como uso por medio de un análisis de los compromisos implícitos de naturaleza deóntica que uno adquiere en el acto mismo de hablar, o pensar, parece encontrar Toribio una fuente de inspiración particularmente caudalosa de cara a articular una idea de esa relación de institución tal que alumbre tanto el modo en que las prácticas sociales instituyen normas que a su vez condicionan esas mismas prácticas como el modo en que nuestras conductas instituyen significados que, de nuevo, condicionan la conducta. Lo que palpita en el fondo de toda esta indagación, que no cabe ya recorrer con mayor detenimiento, es desde luego una voluntad de hacer de la normatividad un rasgo irreductible, pero no inexplicable, de la semántica, y acaso de la semántica, también, un rasgo irreductible pero no inexplicable de lo mental. Habría que ver, con todo, si el camino tanteado por Toribio el replanteamiento de la noción de causalidad que venimos esbozando –no puede decirse, desde luego, que así lo parezca a primera vista.

En todo caso, conviene advertir cómo la cuestión de si es la semántica de los estados mentales, y no su sintaxis, lo que posee determinada soberanía causal parece haber quedado desplazada, en el transcurso del argumento, por la cuestión de si son los estados mentales, y no su encarnación física, los poseedores de dicha autonomía. Más aún: en la controversia sobre el estatus de las leyes *cæteris paribus* en la que se baten Fodor (1991) y Schiffer (1991), y en cuyo seno toma cuerpo la defensa del papel de la semántica que venimos sopesando, queda claro que en el fondo el asunto, tal como lo ve el propio Fodor, ni siquiera atañe particularmente a la eficacia causal de los estados mentales, o a la autonomía de la explicación psicológica, sino que concierne por igual a cualquier otra ciencia que pretenda distinguir su discurso del de la física: "[…] the legitimacy of ceteris paribus laws is an issue not just in the philosophy of psychology but in the philosophy of the special sciences at large" (Fodor 1991: 21). Que la estrategia pasa por fundamentar la eficacia causal de la semántica mediante la aplicación del mismo tipo de argumento –basado en la tesis de realizabilidad múltiple, en la distinción entre instancias y tipos de relación causal, y en la atribución a distintos vocabularios teóricos de la capacidad de captar diferentes generalizaciones sobre tipos de relación causal– que se empleó para justificar la autonomía de las ciencias especiales en general queda explícito en Toribio (1991: 7):

> […t]he thesis that beliefs, desires and other mental states work as internal causes of behaviour and do it so in virtue of their semantic content can be justified by using the same kind of argument that has been used for the special sciences case.

También Jackson y Pettit (1990: 195) anuncian con claridad que, desde su punto de vista, la cuestión de si la semántica de los estados mentales es relevante en la explicación psicológica, como también la de si es preciso atribuirles una estructura sintáctica –y por tanto, el carácter de un *lenguaje* interno– para dar cuenta de las relaciones causales que la concepción funcionalista de lo mental les asigna[348], conforman diversas manifestaciones de un único problema de fondo, el de la eficacia causal de lo mental.

Visto así, no resulta extraño que el cauce de la argumentación siga refluyendo hacia posiciones como las que sostienen precisamente Jackson y Pettit (1990) o Braddon-Mitchell y Jackson (1996), articuladas sobre la distinción entre una relevancia explicativa que se concede a lo mental –o lo semántico, si se prefiere– y una eficacia causal que se le niega. Por distintas rutas hemos regresado ya varias veces a la idea de que esquivar tales conclusiones sólo puede venir dado a hombros de un replantamiento de la noción misma de causalidad. Mientras ese replanteamiento no tiene lugar, mientras sigue dándose por buena de forma más o menos velada una cierta lectura de la idea de herencia de propiedades causales de acuerdo con la cual todas las propiedades causales de una clase de estados mentales

[348] E incluso, al igual que ocurre en Toribio (1991), la de si bajo cierta interpretación de los modelos conexionistas de procesos psicológicos estos contravienen la distinción entre sintaxis y semántica trazada en el seno del cognitivismo clásico, o nos arrastran al eliminacionismo.

son de hecho propiedades causales de los estados físicos en los que se encarnan los estados mentales que forman la clase en cuestión, las aguas acaban siempre por volver al molino del reduccionismo, o del epifenomenismo. Así ocurre, según venimos viendo, en el debate sobre la eficacia causal de lo mental, y así por fuerza habrá de ocurrir también si trasladamos los mismos argumentos a un debate sobre la eficacia causal de la faceta semántica de lo mental, por oposición a su faceta sintáctica –toda vez, además, que se ha definido lo sintáctico en términos que aseguran la replicación de toda propiedad semántica en una sintáctica, y que, *ex definitione*, cifran toda putativa eficacia causal de un estado mental en sus propiedades sintácticas. Esa reflexión sobre las propias ideas de sintaxis y semántica será la última parada de este largo camino.

Nociones de lo sintáctico: pensamiento y lenguaje

La aparente ambivalencia del pensamiento cognitivista ante el valor explicativo que pueda atesorar la semántica de los estados internos incorporados a las teorías psicológicas se entiende mejor, sea como sea, si reparamos en las divergencias entre las nociones de sintaxis y semántica que se ponen en juego en esa reflexión, forjadas como quedó concisamente advertido en la confluencia de la idea de gramática formal de un cálculo lógico y la familiaridad con la mecánica de los primeros autómatas, y las nociones de sintaxis y semántica que hemos heredado de la lingüística, y que siguen operando en el trasfondo de la construcción de teorías sobre cualquier proceso psicológico, acaso incluso sin que se halle involucrado el lenguaje.

Al igual que la idea de que la explicación psicológica pueda arraigar en un nivel de abstracción común a organismos y máquinas –desligado por tanto de la anatomía o la fisiología nerviosa tanto como de la ingeniería eléctrica–, al igual también que la idea de que ese nivel de abstracción alumbra una potente homología entre programas y teorías, o incluso entre programas y procesos psicológicos que dichas teorías procuran explicar (*cf*. *supra*) –de hecho, entrelazándose con ellas a cada paso–, la constatación de que un singularísimo género de vocabulario teórico –el intencional, el teleológico– parecía reclamar su legitimidad para operar en ese nivel de abstracción aun inexplorado fue afianzándose muy paulatinamente a lo largo de las décadas anteriores a la eclosión del cognitivismo. Además del poderoso eco de los trabajos de Hilbert, en efecto, el esfuerzo por diseñar y construir autómatas que pudieran calcar ciertos comportamientos de los seres vivos, y en especial humanos, desempeña un papel insoslayable en esta genealogía de la idea de reivindicar el valor explicativo del léxico conceptual de la psicología engarzando la semántica de lo mental a la sintaxis.

La cuestión de la legitimidad del vocabulario teleológico e intencional, de hecho, comienza, igual que las otras dos, a perfilarse en la controversia biológica a la que Cordeschi ha dado carácter seminal en su visión del "descubrimiento de lo artificial" (Cordeschi 2002: *passim*): el debate entre Jacques Loeb y Herbert Jennings a

cuenta del alcance del tropismo como explicación de conductas complejas –*cf.* Loeb (1900: 240) y Jennings (1910: 354), *supra*. A juicio de Loeb era evidente –y máquinas fototrópicas como la de Hammond y Miessner (*supra*) lo hacían aún más transparente– que la polilla que vuela hacia la luz no lo hacía impulsada por ninguna suerte de curiosidad, querencia, propósito o sentimiento, sino por causas estrictamente mecánicas; también era obvio que lo mismo había de poder afirmarse de cualquier conducta animal o humana, por compleja que fuera: "[…] the analysis of animal conduct", en suma, "only becomes scientific insofar as it drops the question of purpose" (Loeb 1918: 18) –si bien, como recuerda Cordeschi (2002: 9), no la abandona por inviable, sino, al contrario, por haber dado con los mecanismos subyacentes a las conductas que provocaban la ilusión teleológica[349]. Pero nada de aquello –ya lo hemos visto– era tan obvio a ojos de Jennings, quien, más atento que Loeb al hecho de que las similitudes funcionales entre una polilla y el perro eléctrico de Hammond y Miessner convivían con lo que bajo una perspectiva física o química no dejaban de ser abismales diferencias entre ambos (*cf.* Jennings 1910: 361, *supra*), se mostraba convencido de que determinadas generalizaciones que podían expresarse en un nivel de análisis más abstracto se nos escaparían irremediablemente si habíamos de ceñirnos sólo a "la terminología de la física y la química" (Jennings 1910: 354, *supra*). Esas esquivas generalizaciones se remiten –piensa Jennings (1906: 178)– a estados internos del organismo capaces de propiciar distintas respuestas a estímulos físicamente idénticos[350]; aunque dichos estados son descritos siempre como fisiológicos, no como psicológicos, el lenguaje mentalista no tarda en aflorar en la caracterización de la conducta de los organismos a los que se les atribuyen. A tenor únicamente de su utilidad explicativa, de hecho, Jennings no vacilaría en arrogar conciencia –se entiende que lo mismo diría de propósitos o creencias– incluso a organismos a los que raramente lo hacemos:

> If amoeba were so large as to come within our everyday life, I believe beyond question that we should find […] [the] attribution to it of certain states of consciousness a practical

[349] No es difícil advertir en Loeb, así pues, unos primeros cabeceos de la fluctuación entre afanes reduccionistas y eliminacionistas que acabaría por configurar, como venimos viendo, un rasgo tan común en la fisonomía conceptual tanto del conductismo como del fisicalismo de orientación neurofisiológica.

[350] Es palpable, así pues, el engarce entre el papel que Jennings asigna a los estados internos y la tesis que Thorndike trataría de instaurar no mucho después como "un fragmento de la ley general de uniformidad de la naturaleza" al que bautizó como *primera ley de la conducta*: que si una misma situación produce en el mismo animal respuestas diferentes en ocasiones diferentes, el estado del animal debe haber cambiado, pues si el estado del animal es idéntico, idéntica situación provocará idéntica respuesta (Thorndike 1911: 241).

El manantial del que fluyen las intuiciones cardinales del internismo y el solipsismo metodológico –*cf.* Guttenplan (1994b, *supra*)– no parece, después de tanto litigio entre conductismo y cognitivismo, encontrarse muy apartado de estos parajes: sólo cabe esperar distinto comportamiento ante idéntica situación cuando medien distintos estados internos; son esos estados internos –en particular, la representación de la situación que se forma el organismo más que la propia situación– lo que procede incorporar a la explicación de su comportamiento.

> assistance in foreseeing and controlling its behavior […]. In a small way this is still true for the investigator who wishes to appreciate and predict the behavior of the amoeba under microscope. (Jennings 1906: 337)

Detenerse un instante en el examen de los planteamientos de Jennings permite reparar en la debilidad del vínculo entre un lenguaje mentalista predicado globalmente del organismo, por una parte, y, por otra, ciertos estados internos cuya mediación funcional entre estímulos y respuestas hace la explicación y predicción de la conducta del organismo inasequible al vocabulario de la física y la química, pero de los que no se predican con claridad propiedades intencionales o teleológicas. Eso sí: como en el cognitivismo posterior, que a esos estados internos no se les atribuya, en su imbricación en la conducta del organismo, incompatibilidad alguna con los principios explicativos de la física y la química, no es óbice para mantener que tal imbricación hace verdaderas generalizaciones sobre la conducta del organismo que escapan a dichos principios (*cf*. Cordeschi 2002: 20) –en el caso de Jennings, sin embargo, es la plausibilidad de una reducción de la fisiología a la física, no de la psicología a la fisiología, lo que está en entredicho.

Animado acaso por la polémica entre Loeb y Jennings, así como por su conocimiento de los primeros aparatos manufacturados con ánimo de imitación de la vida, un discípulo de William James y destacado adversario del idealismo como Ralph B. Perry alcanzaba ya en 1917 a plantearse con toda claridad en qué medida el vocabulario teleológico que podamos emplear en la explicación de la conducta de organismos o máquinas es imprescindible para que dicha explicación sea efectiva, o resulta más bien algo así como una perspectiva –lo que desde Dennett (1971, 1987) nos hemos acostumbrado a llamar "a stance", una *actitud*– que el observador puede adoptar o desechar sin mayores consecuencias en lo que atañe a –digamos– su solvencia epistemológica[351]. Es más, Perry repara con toda nitidez en que admitir que el vocabulario teleológico resulte imprescindible entraña asumir que existe "[…] algún factor adicional" (Perry 1917: 359 *apud* Cordeschi 2002: 133) que haga insuficientes las explicaciones íntegramente mecánicas o neurofisiológicas, así como en la dificultad de dar cuenta de la naturaleza de dicho factor sin convertirlo en un

[351] La idea de que la atribución de estados mentales a un organismo o una máquina, o de propósitos a sus comportamientos, no tenga otro fundamento que la adopción pragmática de cierta perspectiva se remonta, si no a Jennings (1906: 337, supra), sí al menos hasta Rosenblueth, Wiener y Bigelow (1943). Así ha sabido verlo también Cordeschi (2002: 259-260), quien recuerda además como MacKay (1962) caracterizaba la adopción de una "actitud personal" hacia el organismo o la máquina cuya conducta tratamos de entender como una decisión del observador. Ya el propio MacKay (1952) había *decidido* describir la relación entre los estímulos y respuestas de sistemas retroalimentados –termostatos, por ejemplo– como mediadas por una representación interna, un "[…] correlato simbólico" (MacKay 1952: 73) al que la respuesta tiende a equiparar el estímulo, y había sugerido que tal mecanismo bien podía constituir la base de todas las funciones psicólogicas superiores.

Sean cuales sean sus otras virtudes, parece claro a estas alturas que sobre una posición instrumentalista de esta índole no puede descansar ninguna idea de la explicación psicológica que la haga irreductible a explicaciones expresadas en vocabularios más escuetos, como el de la neurofisiología o el de la física –como no sea, *i.e.*, irreductible por motivos meramente pragmáticos.

ente poco menos que milagroso –sin revocar el naturalismo de la explicación. La relevancia del tipo de vocabulario teórico que emplee una determinada disciplina resulta capital a ojos de Perry, que en su matizada defensa del conductismo señala como uno de los principales méritos de éste haber aclarado que la diferencia entre fisiología y psicología no puede ser de objeto de estudio –"[…] como la diferencia entre la entomología y la ornitología" (Perry 1921: 85)– sino de enfoque, una constatación que él considera, también, de raigambre aristotélica[352].

El intento de Perry de conferir a la idea de propósito una lectura estrictamente objetiva mereció la estima de Tolman, que de hecho se mostraba menos exigente que el propio Perry a la hora de admitir que la apelación al propósito era ineludible en la explicación de una conducta dada –lo que ocurría, a su juicio, siempre que la descripción de la conducta en cuestión debiera aludir a propiedades de aquello que constituía su objetivo (Tolman 1925: 37). Parecida discrepancia se da entre Tolman y Hull: si bien, como se ha visto, Hull (1930: 514, *supra*) admitía que el mundo imprime sobre el organismo una representación de sí mismo que contribuye al control de la conducta, sería a través del concepto de *mapa cognitivo* articulado por Tolman (1948) como tal representación quedaría expresamente ligada al carácter teleológico de la conducta –un mapa cognitivo representa el entorno en relación con los fines del organismo y los medios de los que dispone para alcanzarlos–, y de paso se convertiría en el mecanismo fundamental para explicar el aprendizaje, en detrimento del reforzamiento o la debilitación de conexiones entre estímulos y respuestas. En círculos cercanos a Hull, no obstante, se desplegó una vigorosa defensa del papel del vocabulario teleológico e intencional en la explicación de ciertas conductas: Nicolas Rashevsky (1931) reparó en que determinados procesos de aprendizaje que podemos atribuir tanto a organismos como a autómatas tienen lugar "[…] on a different level from that of the ordinary muscular reaction" (Rashevsky 1931: 393 *apud* Cordeschi 2002: 98), lo cual lo condujo a la convicción de que el empleo de conceptos mentalistas era ineludible si habíamos de construir una teoría general de la conducta de esos distintos sistemas físicos.

La idea de representación interna como sustrato común a la explicación de la conducta de los organismos y la de ciertas máquinas iba también madurando en ámbitos más apartados de la psicología. El influyente ensayo sobre la aplicación de los principios matemáticos de la termodinámica al estudio de la biología, y en

[352] Es indudable que la observación de Perry anticipa algunos aspectos cardinales de la interpretación funcionalista de la autonomía de la explicación psicológica. Su tesis, sin embargo, es que la diferencia entre psicología y fisiología es, *in nuce*, la que media entre una aproximación molar y una molecular a los determinantes de la conducta, idea que el cognitivista rechazaría –junto, claro está, con toda restricción a la incorporación a nuestras explicaciones de conceptos referidos a estados o procesos internos– para reemplazarla por la de diferencias en cuanto al nivel de abstracción.

El acercamiento de Perry a la convivencia entre el saber psicológico y el fisiológico parece haber dejado una honda huella en el conductismo: Orval H. Mowrer (1960), entre tantos otros, viene a defender la misma idea, cuyo ímpetu retrotrae hasta las dudas de Thorndike (1931) respecto a que la neurona pudiera ser la unidad de análisis apropiada en lo que atañe a la explicación de la conducta humana –*cf*. Cordeschi (2002: 56).

especial de la ecología de poblaciones, que Alfred J. Lotka presentó en 1925 incluía la descripción de un sencillo autómata, una especie de escarabajo mecánico capaz de desplazarse sobre una mesa sin caerse gracias a una antena que detectaba los bordes. De acuerdo con la interpretación del propio Lotka, el estado de la antena "[…] depicts, in a crude but sufficient manner the environment *in* the toy" (Lotka 1925: 341-342, énfasis añadido). La idea de que los seres humanos, como otros muchos animales, poseemos un aparato representacional de características parecidas surge reiteradamente en las reflexiones de Lotka –su expresión más concisa es quizá la tesis de que "[…t]he external world is depicted in the organism by a certain apparatus, a set of organs and faculties, which we may appropriately term the Depictors" (Lotka 1925: 339)–, si bien la idea de representación que Lotka maneja –ingenua, si se quiere– se perfila como equivalente a la de correlación con determinadas propiedades del entorno[353].

Es sin embargo en Craik (1943) donde podemos encontrar el primer esfuerzo sistemático de reflexión acerca de aparatos representacionales como los postulados por Lotka (1925), Hull (1930) o Rashevsky (1931) –atendiendo a esa primacía ha acuñado Johnson-Laird (1983: 400) la expresión "autómata craikiano" para referirse, por contraposición a los "autómatas cartesianos", a máquinas que cuentan con cierta capacidad de representación de su entorno. En *The Nature of Explanation* no tenemos ya sólo estados físicos del organismo o del artefacto que correlacionan con propiedades ambientales y que son capaces de guiar la conducta, sino también un aparato inferencial que por medio de operaciones simbólicas dota al sistema de la facultad de extraer consecuencias de su representación del entorno que no vienen directamente causadas por éste a través de los mecanismos receptores[354].

Con el alborear de la computación digital, la intuición de que rendir cuenta de ciertas conductas, ya fuera en organismos o en autómatas, podía exigirnos adoptar un nivel de abstracción superior al que podía proporcionarnos el vocabulario de la física o la neurofisiología se convirtió en una constatación práctica e inmediata: la enorme dificultad que revistía, incluso tratándose de tareas mínimamente complejas, la programación en código-máquina –donde las operaciones se describen en términos de posiciones literales de conmutadores físicos– hizo perentorio el

[353] Exagera tal vez Cordeschi (2002: 267) al afirmar que Lotka concibe ya ciertos estados del organismo o la máquina como "[…] symbols denoting world states", en la medida en que la noción de símbolo parece exigir pertenencia a un sistema simbólico de construcción de significados y no mera correlación con hechos o propiedades. Sea como sea, un lúcido examen de la relevancia de las conclusiones de Lotka en lo concerniente a la concepción de la explicación teleológica que promovería el avance de la cibernética puede hallarse en Cordeshi (2002: 127-128).

[354] Es decir, que en un autómata craikiano sí que podríamos hablar con propiedad de símbolos, pues la imbricación en procesos inferenciales constituye exactamente esa pertenencia a un sistema simbólico de construcción de significados de la que adolecía el escarabajo de Lotka –al que, como es natural, Cordeschi (2002: 138), ajeno a ese matiz, considera un autómata craikiano *ante litteram*. En todo caso, la definición de autómata craikiano en Johnson-Laird (1983) es suficientemente lata como para dar cabida a ambas interpretaciones.

desarrollo de lenguajes de alto nivel[355]. Ya en la conferencia de Darmouth College, aquel verano de 1956, el hecho de que Newell, Shaw y Simon no hubieran escrito en código-máquina el programa de su *Téorico Lógico* fue motivo del penetrante interés de John McCarthy –quien luego, como es sabido, desarrollaría su propio lenguaje de alto nivel, el LISP [LIS*t* P*rocessor*]. Pero no ya la programación: ni siquiera la interpretación del lenguaje de máquina era, las más de las veces, asequible a los propios ingenieros y programadores –con la expresión que años después haría célebre Douglas Hofstatder (1979: 290), "[…l]ooking at a program written in machine language is vaguely comparable to looking at a DNA molecule atom by atom". Además de simplificar el día a día de los programadores, los lenguajes de alto nivel tenían la importante ventaja de permitir la portabilidad del programa entre máquinas con arquitecturas físicas diferentes: exactamente, *mutatis mutandis*, aquello a lo que venían apuntando, como se ha visto, distintas reflexiones sobre la necesidad de un vocabulario explicativo arraigado en un nivel de abstracción de índole funcional, en el que sistemas nerviosos e ingenios hidráulicos o electrónicos compartieran un mismo nicho. Que dicho vocabulario explicativo hubiera de caracterizarse por redimir nociones *subjetivas* como significado o propósito era, desde luego, otra cuestión; ahora bien, cuando se trataba de la simulación computacional de ciertas conductas humanas, donde, como se ha visto, el propio programa se iba perfilando como teoría de los procesos psicológicos involucrados o incluso como instanciación de dichos procesos, había de ir cobrando un vigor creciente la idea de que el lenguaje de programación idóneo y, con ello, el lenguaje teórico idóneo compartieran con el vocabulario psicológico tradicional determinados rasgos cardinales, como, precisamente, su carga teleológica o intencional. En lo que atañe a Newell, Shaw y Simon, su apuesta era tajante: "The invention of the digital computer […] has provided us with operational and unobjectionable interpretations of terms like 'purpose,' 'set' and 'insight'" (Newell, Shaw y Simon 1958: 162). Diez años después, Marvin Minsky (1969: 2) podía ya aducir el uso de vocabulario mentalista como prueba de la distancia que separaba a la investigación en inteligencia artificial del conductismo. No es menos cierto, sin embargo, como ha hecho ver Dennett (1986), que al margen de lo que pudiera proclamarse *de iure* respecto al estatus del vocabulario mentalista, en el intento pionero de Newell y Simon (1972, 1975) de explicar los procesos de solución de problemas como el paso guiado por reglas formalmente especificadas de una representación formal del problema a una representación formal de la solución –igual que en el afán de McCarthy (1980) por desarrollar una representación formal del conocimiento necesario para probar la adecuación de varios planes de acción a unas circunstancias dadas–, se daba por

[355] Entre el código-máquina y los lenguajes de alto nivel, en rigor, median lenguajes de bajo nivel, como el ensamblador, que aún exigen al programador un conocimiento detallado de la arquitectura física de la computadora. La teoría de los lenguajes de programación es un territorio cuya fertilidad de cara a la comprensión de la naturaleza de la explicación psicológica no es posible abordar más detenidamente en este trabajo.

buena *de facto* la condición de formalidad estipulada por Fodor –en efecto, *cf*. en este sentido Newell, Shaw y Simon (1957: 129) o Simon y Siklóssy (1972: 2), *supra*.

Por otra parte, una intensa carga semántica impregna nuestras ideas comunes de categoría sintáctica, articuladas en torno a las de *sujeto* y *predicado* –como impregna también todos los otros niveles de estudio del lenguaje que solemos distinguir, excepción hecha del estrictamente fonético o del análisis de glifos que le correspondería en el ámbito de la lengua escrita. Es bien sabido que cuando menos desde que el trabajo de Antoine Arnauld y Claude Lancelot, continuado luego por sus discípulos de Port-Royal-des-Champs, fructificara en la *Grammaire générale et raisonnée*, se procuró dar a esas categorías una universalidad que las hiciera aplicables a cualquier lenguaje natural, y que había de provenir de su imbricación en categorías lógicas que se entendían como consustanciales al pensamiento que el lenguaje expresaba. El intento de liberar la sintaxis de servidumbres semánticas, convirtiéndola en una combinatoria estrictamente formal que nos provea de las reglas sintagmáticas y paradigmáticas según las cuales determinados constituyentes conforman construcciones oracionales o suboracionales válidas, es –qué duda cabe– uno de las fuerzas mayores que han impulsado la lingüística en los últimos tiempos, y de hecho expresiones tan medulares de ese proyecto como la gramática generativa han mantenido estrechísimas relaciones, como es bien sabido, con el cognitivismo en psicología.

Sin embargo, no es descabellado argumentar que el intento de modelar toda propiedad semántica bajo una regimentación puramente sintáctica, en el sentido de formal, desborda no ya sólo la idea tradicional de categoría sintáctica, sino incluso los objetivos de las primeras gramáticas inspiradas en los trabajos seminales de Chomsky (1957). La razón radica, fundamentalmente, en que los sistemas formales de la gramática chomskiana aspiran a generar clases de expresiones identificables con las que los hablantes de una lengua reconocerían como gramaticales, pero no a condensar cualquier distinción semántica que dichos hablantes pudieran perfilar. Antes bien, la propia tesis de autonomía de la sintaxis alcanza buena parte de su vigor a partir del conocido contraste que Chomsky (1966: 35, *cf*. Chomsky 1957: 15) estableciera entre "Revolutionary new ideas appear infrequently" y "Colorless green ideas sleep furiously", dos enunciados entre los que media una disparidad semántica que difícilmente podría ser reconstruida en términos sintácticos, siendo así que las estructuras sintácticas de ambos son a duras penas discernibles. No cabe identificar, pues, estructura y significado –lo que tanto conviene a una defensa de la autonomía de la sintaxis como de la irreductibilidad de la semántica.

Es innegable, en todo caso, que el proyecto de Fodor –el de afinar nuestras distinciones sintácticas de suerte que toda diferencia semántica descanse en una– no carece de parangón en el ámbito lingüístico, en especial en el ámbito del programa minimalista alentado por el propio Chomsky (1993): una reducción o una drástica eliminación de la semántica es la encrucijada a la que Uriagereka (2008), por ejemplo, considera que nos aboca su transparente colinearidad con la sintaxis, que se desvela tan pronto como vamos disponiendo de herramientas de análisis suficientemente

afinadas. Hay también, desde luego, proyectos de signo opuesto, como el de fundamentar una lingüística que se reclama genuinamente cognitiva –y que reconoce también sus deudas con la semántica generativa: *cf*. Cuenca y Hilferty (1999: 20)– sobre lo que Lakoff y Johnson (1987) denominaron realismo experiencial, que no es sino un abierto cuestionamiento de la tesis de que el pensamiento –mucho menos, claro, la totalidad de la vida mental– pueda, dada su inextricable imbricación en la experiencia de nuestro propio cuerpo, el mundo y, particularmente, nuestro entorno social, entenderse cabalmente como un complejo mecanismo de manipulación de representaciones. Lo que se plantea entonces es, como en Langacker (1987: 35), "[…] un *continuum* de unidades simbólicas" formado por el léxico, la morfología y la sintaxis, "[…] que sirven para estructurar el contenido conceptual con finalidades expresivas" y hacen "[…] incoherente hablar de la gramática separada del significado", pues todas y cada una de esas unidades simbólicas se articulan entre un polo fonológico y un polo semántico. Es decir, en suma, que "[…] la semántica incide en la gramática: la gramática es un vehículo de la semántica y, como tal, la sintaxis se ve 'contaminada' por ella" (Cuenca y Hilferty 1999: 94-95). Naturalmente, una tasación siquiera tentativa de hasta qué punto cada uno de estos proyectos enfrentados es hacedero requeriría un despliegue mucho más minucioso de lo que refrendan los límites de esta investigación. Aun así, conviene siquiera tantear el modo en que la persistencia de nociones tradicionales de lo sintáctico y lo semántico, o bien la reivindicación de los lazos entre ambas, dan forma, entremezcladas con otras de inspiración más netamente formalista, al debate sobre el papel de la semántica en la explicación psicológica.

En el plano del lenguaje natural, desde luego, es fácil encontrar diferencias semánticas que –como en "Colorless green ideas…"– no parecen estar incrustadas en la sintaxis, al menos ciñéndonos a aquel sentido tradicional de la distinción entre lo semántico y lo sintáctico al que venimos aludiendo. Cualquier par de oraciones en las que un término polisémico cumpla idéntica función sintáctica pero aporte al significado de cada oración distinto valor semántico –o en el que otro tanto ocurra con dos términos homónimos– basta para construir un ejemplo. Del mismo modo, no es difícil dar con construcciones sintácticamente ambiguas en las que la interpretación semántica de la oración quede comprometida. Un caso clásico, que Quine (1960: 177) tomó de Peirce (1932) y éste de Allen y Greenough, es el siguiente: "Un abogado dijo a un colega que creía que uno de sus clientes le criticaba más que a sus rivales". ¿Cuál es la función sintáctica del sintagma "a sus rivales"? Como parte del segundo término de una estructura comparativa, desempeña la función de baremo adjunto al transpositor "que" y ligado a un verbo elidido, "[criticaba]", por coincidir con la base de la comparación explícita en el primer término. Correcto. ¿Y cuál es la función del pronombre posesivo "sus" en el marco del sintagma? Más exactamente: ¿cuál es su antecedente? ¿Se trata de los rivales del abogado, de los de su colega, o de los del cliente? El análisis sintáctico, obviamente, no tiene respuesta para esta pregunta: se trata de una ambigüedad sintáctica. Por supuesto, la ambigüedad infecta la semántica de la frase –lo extraño sería dar con un caso de

ambigüedad sintáctica en que esto no sucediera–, así que no sabemos cuál es exactamente el hecho que ésta describe. Pero eso no la convierte en una ambigüedad semántica, porque su origen no se halla en que el adjetivo posesivo oscile entre varios significados posibles –como ocurre debido a la polisemia de "banco" en "Llegamos al banco que está junto al parque"–, sino entre varias construcciones sintácticas[356].

Pero no es en el lenguaje natural donde andábamos buscando fenómenos de esta índole. Lo que buscamos son diferencias en las propiedades semánticas de la *lingua mentis* que –supongamos– comparten dos sujetos, no en las de sus lenguajes naturales –diferencias, pues, que habrán de encarnarse en señales nerviosas, no en fonemas o grafemas–, y que puedan subyacer a una diferencia conductual entre ambos, obligándonos así a tomarlas en cuenta a la hora de predecir sus conductas o relatar el itinerario de causas y efectos que las ha originado. La pregunta es, entonces: si tomásemos una oración como la de Allen y Greenough como expresión de una creencia, ¿qué tendríamos que encontrarnos para que fuera razonable concluir que su ambigüedad sintáctica no se debe, en un caso particular, a un deficiente funcionamiento de los mecanismos de producción del habla –que, por así decir, estarían expresando de manera confusa una idea clara–, sino a que la ambigüedad está presente en la propia representación mental de la creencia –o sea, a que la idea, y no sólo su expresión, es confusa?

Parece razonable convenir en que si el sujeto que expresa la creencia diera su asentimiento a ésta bajo una de las interpretaciones posibles y actuara en consecuencia, sin advertir o incluso rechazando abiertamente la existencia de otras interpretaciones, tendríamos que inclinarnos por la conclusión de que no hay ambigüedad en su representación mental de la creencia, y de que la ambigüedad

---

[356] Sin embargo, la relación del adjetivo posesivo con su antecedente no es aquí de orden morfosintáctico, puesto que la única marca morfológica de "sus", la de plural, obedece a la concordancia con el sustantivo que designa lo poseído, "rivales", no con el antecedente del posesivo, cualquiera que éste sea. Así es preceptivo en castellano, que no diferencia la tercera persona del singular de la del plural en sus posesivos. Así, la ambigüedad no se resolvería ni siquiera en "Un abogado dijo a un colega que creía que sus clientes le criticaban más que a sus rivales", que la harían difuminarse parcialmente en lenguas como el inglés, en que el posesivo concuerda con su antecedente. Que una ambigüedad como esta resulte en algunas lenguas refractaria al análisis morfosintáctico nos muestra que su fuente mana más bien de territorio sintáctico-semántico.

Si tradicionalmente la función sintáctica de sujeto se especificaba mediante claves semánticas, como la de que el sujeto (para una oración activa) designa a quien lleva a cabo la acción del verbo, algo semejante sucede con los posesivos. En efecto, la función sintáctica de un adjetivo posesivo parece ser la de trasladar a otro contexto oracional al constituyente de la oración que denominamos "antecedente", y que es aquél que desempeña la función semántica de designar al poseedor del objeto con el que concuerda el adjetivo, o bien, con frecuencia, injertar en un determinado contexto oracional una referencia lingüística a algún aspecto de la realidad extralingüística que oyente y hablante comparten pragmáticamente: en "Dale su lápiz", por ejemplo, "su" injerta en la oración imperativa una referencia lingüística al poseedor del lápiz, el cual no ha sido mencionado previamente porque –suponemos– es pragmáticamente innecesario hacerlo. Aislar, en los lenguajes naturales, la sintaxis de la semántica es sin duda una labor ardua, si es que es factible.

presente en la oración con la cual la expresa se ha originado al verter su creencia al habla sin que –por cierto– el sujeto tenga consciencia de ello. Si, por el contrario, el sujeto diera su asentimiento a una sola de las interpretaciones, pero reconociera la gramaticalidad y el significado de las otras, la conclusión no cambiaría salvo en lo relativo a su nivel de competencia metalingüística. En cambio, si el sujeto asintiera a más de una interpretación de la frase con que expresa su creencia, ya lo hiciera –por así decir– de pensamiento, palabra, obra u omisión, no nos quedaría otro remedio que asumir que su representación mental de la creencia es tan ambigua como lo es su expresión lingüística de ésta. Podríamos encontrarnos –por ejemplo– con que sujetos que han sido expuestos a la frase tendieran a responder positivamente tanto a la pregunta de si el cliente era más crítico con el abogado que con los clientes del abogado, como a la de si era más crítico con el abogado que con los clientes del colega del abogado, o a cualquiera de las múltiples combinaciones que permiten las varias ambigüedades encerradas en la oración de Allen y Greenough. Si además encontráramos que este efecto no se da en sujetos que han sido expuestos a una versión minuciosamente desambiguada de la oración, como "Un abogado dijo a un colega que uno de los clientes del colega era más crítico con el colega que con los rivales del colega" (o, de manera mucho menos artificiosa merced al estilo directo: "Un abogado dijo a un colega: 'Uno de tus clientes es más crítico contigo que con tus rivales'"), tendríamos buenas razones para concluir –con todas las salvaguardas metodológicas– que en los sujetos expuestos a la condición experimental hemos logrado implantar representaciones mentales sintácticamente ambiguas. Por lo demás, no parece que un resultado experimental como el descrito fuera particularmente insólito, sobre todo si entre la exposición al estímulo y la ejecución de la tarea media un tiempo que baste para que la representación se descomponga y recomponga en la memoria. Si esa intuición es acertada, tampoco sería aventurado admitir que algunas anomalías sintácticas serán al cabo atribuibles a fallos de producción del lenguaje –fallos tardíos, digamos–, y otras a la sintaxis de la representación mental subyacente –fallos tempranos. No podemos transitar sin más de la anomalía sintáctica en el plano lingüístico a la determinación de la conducta, porque el paso por la sintaxis de la representación mental no es un paso franco.

No está claro, sin embargo, que suceda lo mismo en el ámbito de las anomalías semánticas que pueblan el reino de lo intensional. A diferencia de lo que ocurre con la sintaxis, al hablar de las propiedades semánticas de un estado psicológico y de las propiedades semánticas de la oración que lo expresa o lo describe partimos de un presupuesto de identidad entre ellas. Si mi creencia de que *Fab* tiene la propiedad semántica de referirse a un objeto *a*, del cual predica que mantiene una relación *F* con otro objeto, *b*, entonces –asumimos– la oración que designa mi creencia heredará esas mismas propiedades semánticas, por muy distinta que su estructura sintáctica pueda ser de la que exhiba la oración con la que exprese lingüísticamente tal creencia. De ahí, por ejemplo, que cuando la oración que enuncia una creencia se muestra referencialmente opaca, demos por bueno que la representación mental de la creencia será también opaca. Ahora bien, si abandonamos el plano de la expresión de la

creencia para transitar hacia el de su función psicológica, lo que encontraremos es, desde luego, que si yo creo que *Fab*, actuaré como actuaría si realmente fuese el caso de que *Fab* –haya o no *a*, mantenga *a* la relación *F* o no con *b*, sea o no *a* idéntico a *c*, de quien acaso crea que no es el caso de que *Fcb*, etc. Ésta es, precisamente, la intuición solipsista (*cf*. Guttenplan 1994b: 290) respecto a la explicación de la conducta, que –por lo que vemos– tanto puede servir para minar la credibilidad de la tesis de que dicha explicación pueda pasar sin aludir a la representación que un organismo o un autómata se labren de su entorno (*cf*. *inter alia* Miller, Galanter y Pribram 1960: 17 o Pylyshyn 1984: 6, 216, *supra*), como –según comenzamos a vislumbrar ahora– para estrechar de tal forma los lazos entre la semántica de la expresión lingüística de una actitud proposicional y su hipotética instanciación en el lenguaje del pensamiento que la semántica de aquella resulte superflua tan pronto como la de ésta haya quedado del todo sometida a la tutela de una sintaxis estrictamente formal. Así pues, la intensionalidad de las expresiones lingüísticas de nuestras actitudes proposicionales y la notoria refractariedad a los hechos del papel de dichas actitudes en nuestra economía cognitiva –que son el mismo fenómeno, contemplado ya desde una atalaya lingüística, ya psicológica– parecen garantizar la ineficacia causal de la semántica y, con ello, la prosperidad del patrón de explicación solipsista. Cuando se insiste, entonces, en que la semántica de un estado mental, a diferencia de su sintaxis, no es apta para inmiscuirse en las concatenaciones de causas y efectos que acaban provocando una u otra conducta debido a su carácter *relacional*, el vigor de la argumentación reside en realidad en la idea de que el objeto de una relación *intencional* no puede constituir una causa, precisamente debido a las singularidades de la intencionalidad como relación: de otro modo, el mismo argumento serviría para vetar toda eficacia causal también a la sintaxis, que no atañe, después de todo, sino a *relaciones* entre símbolos –propiedades, pues, tan poco *locales* como las semánticas, aunque, eso sí, no de naturaleza intencional.

Esta constatación nos permite contemplar la polémica en torno al papel de la semántica en la explicación psicológica desde un ángulo diferente. En el núcleo del cognitivismo habita –bien lo sabemos– una rotunda reivindicación del significado, o de algo que cabe llamar así, como elemento indispensable en la comprensión de la conducta. Frente al conductismo, los primeros cognitivistas insistían –así por ejemplo Miller, Galanter y Pribram (1960: 17, 23, *supra*)– en que no son al fin y al cabo los estímulos, sino la representación de los estímulos que se forja el sujeto lo que determina su conducta: un modo de pensar esta idea era que lo decisivo en psicología había de ser precisamente el significado, lo que las cosas significan para uno. Sin embargo, la articulación metateórica de esta misma clave –su ensamblaje en el armazón que proporcionaba la lectura funcionalista de la teoría de autómatas– desembocaría en la tesis de Fodor (1980a: 64, *supra*) según la cual la semántica de los estados mentales –es decir, las cosas o los hechos con las que establecen lazos semánticos de referencia o de verdad– parece mostrarse inerte en el

desenvolvimiento de la vida mental, y desde luego en la determinación de la conducta[357].

La apariencia de contradicción es intensa, pero es fácil hacer traslucir que no hay contradicción alguna en aseverar que no es *lo que significan nuestros pensamientos* –esto es, las cosas a las que se refieren, los estímulos– lo que guía nuestra conducta, sino *lo que las cosas significan para nosotros* –esto es, la representación de las cosas, de los estímulos si se prefiere, en nuestros pensamientos, no los estímulos, o las cosas representadas, en sí. Porque lo que las cosas significan para nosotros puede entenderse como su representación mental, y lo que significan nuestros pensamientos puede entenderse como las propias cosas, una velada *equivocatio* sobre el significado de "significado" –por tomar prestada la expresión que Ogden y Richards (1923), y luego Putnam (1975a), acuñaron con distintos propósitos– subyace a la ambivalencia con que el cognitivismo se enfrenta al valor de la semántica en la explicación psicológica. Dicho de otra manera: aunque parecen a simple vista dos gestos contradictorios, la reivindicación del significado –el modo en que nos representamos las cosas– como fundamento de la explicación psicológica y la insistencia en que el significado –las cosas que de ese modo nos representamos– es inerte en la determinación de la conducta que tratamos de explicar son, a fin de cuentas, uno solo y el mismo gesto.

Bajo este prisma, que las formulaciones más tempranas del funcionalismo, al abrigo de la teoría computacional de la mente, se inclinaran por una posición internista acerca de la semántica de los estados mentales aparece como la desembocadura natural de las corrientes históricas, pues el funcionalismo fructifica precisamente, en buena medida, como una reacción ante la tendencia de los conductistas a prescindir del eslabón interno en la determinación de la conducta. En efecto, la convicción que motivaría a muchos psicólogos a quebrantar la estricta regimentación de la actividad teórica y experimental impuesta por el canon conductista sería, precisamente, que un concepto de estímulo capaz de llevar el peso de la explicación de la conducta parecería aludir más a la representación de su entorno por parte del organismo que a dicho entorno *per se*. Tal como Graham (2007) describe la primera de las razones que revisa para no adherirse al conductismo:

> The fact that the environment is represented by me, to me, constrains or informs the functional relations that hold between my behavior and the environment and may, from an anti-behaviorist perspective, partially disengage my behavior from its reinforcement history. No matter, for example, how tirelessly and repeatedly I have been reinforced for pointing to or eating ice cream, such a history is impotent if I just don't see a potential stimulus as ice cream or represent it to myself as ice cream or if I desire to hide the fact that something is ice cream from others. (Graham 2007: §7)

Pero son precisamente esta clase de razonamientos, como hemos visto, los que conforman las intuiciones básicas que, en relación con la explicación de la acción,

[357] Aunque *cf*. Fodor (1989: 67, *supra*).

alientan la concepción internista de la semántica de lo mental –*cf.*, por ejemplo, Guttenplan (1994b: 290, *supra*). Es del todo entendible, así pues, que, en plena debacle de la concepción conductista de lo mental, la prometedora alternativa funcionalista se desplegara bajo un marcado sesgo internista y que, con ello, el papel de la semántica en la explicación psicológica quedara irónicamente empañado.

Pero junto a todas estas –digamos– buenas razones que parecen alejarnos del hallazgo de diferencias semánticas conductualmente relevantes entre representaciones internas conviven otras que tienen que ver con el modo en que la distinción entre sintaxis y semántica se ha diseminado en el ámbito de las ciencias cognitivas, y que acaso sean en buena medida espurias. Es obvio que una vez hayamos abstraído como propiedades sintácticas de los estados de un sistema aquellas de entre sus propiedades físicas que resulten relevantes de cara al control de la conducta por razón de sus conexiones con otros estados del sistema, incluidos los de sus aparatos de naturaleza perceptiva y motora –pero no, por ejemplo, los que atañen al mero mantenimiento de la actividad eléctrica–, cualquier otra propiedad, incluida cualquier referencia a lo que dichos estados del sistema puedan representar o significar, será prescindible a efectos de predecir la conducta o de rendir cuentas del proceso exacto que la ha producido[358]. Sabemos, en efecto, que no son propiedades sintácticas todas las propiedades físicas de la actividad nerviosa, ya que "[…] toda caracterización sintáctica supone una considerable abstracción respecto de las propiedades físicas" (Dennett 1982: 21)[359], sino sólo aquellas que intervienen en el desarrollo de funciones cognitivas y conductuales. La imbricación entre lo sintáctico y lo funcionalmente eficaz se refleja, por ejemplo, en la manera en que Dennett se deshace de las inscripciones que adornan –o describen para nosotros– los nodos de una red semántica, que evoca poderosamente la inocuidad de los números y signos que se grababan sobre las ruedas de la máquina de Babbage, *supra*: "[…l]a verdadera 'sintaxis', la estructuración del sistema *de la que depende su funcionamiento*, está toda en los vínculos" (Dennett 1981: 31, énfasis añadido). La interpretación natural de la cláusula que aparece entre comas es que se trata de una aposición explicativa que expresa la identidad entre lo denotado por el antecedente y lo denotado por la propia aposición: es decir, que la sintaxis *es* "[…] la estructuración del sistema de la que depende su funcionamiento".

Idéntica noción de la sintaxis trasluce en una analogía que el propio Dennett (1982) emplearía poco después para remachar la tesis de la ineficacia de la semántica, y que recuerda a las consideraciones de Cummins (1989: 84, *supra*) sobre la insensibilidad de las representaciones internas postuladas por la teoría

[358] Con todo, aun si el conocimiento de las propiedades sintácticas del sistema, así definidas, nos asegurase la capacidad de predecir primero y rastrear después el origen causal de cada una de sus conductas, ello no entraña que nos asegurara también la capacidad de explicarlas –o de entenderlas. La insistencia de Pylyshyn (1984: *passim*, *supra*) en diferenciar descripción de explicación vuelve a cobrar relieve en este contexto.

[359] *Idem*, recuérdese, en Fodor (1985: 26, *supra*): "To a first approximation, we can think of [a symbol's] syntactic structure as an abstract feature of its (geometric or acoustic) *shape*".

computacional de la mente a propiedades históricas de sus referentes[360]. De la misma manera –arguye Dennett– que ningún dispositivo físico podría detectar cuál de dos monedas ha estado diez minutos sobre mi mesa, tampoco podría detectar cuál de dos creencias es acerca de un objeto determinado: ni el tiempo que un objeto ha permanecido en determinado estado o lugar es una propiedad física, o al menos una detectable –se infiere–, ni lo es el contenido semántico de una creencia. Pero si, por la razón que fuera, la permanencia de la moneda sobre la mesa hubiera dejado alguna huella física, entonces habría de existir algún dispositivo físico capaz de detectarla, igual que sucedería si el hecho de que una creencia se refiriese a un objeto determinado viniera ligado a propiedades físicas detectables. Dado que ningún dispositivo físico puede ser sensible al contenido semántico de las creencias, y que el cerebro es un dispositivo físico, se sigue que el cerebro es insensible al contenido semántico[361]. Así pues:

> The brain's testing of *semantic* properties of signals and states in the nervous system must be similarly indirect testing, driven by merely syntactic properties of the items being discriminated –that is, by whatever structural properties the items have that are amenable to direct mechanical test. (Dennett 1982: 26)

Propiedades sintácticas de una representación interna parecen ser entonces, a ojos de Dennett, aquellas propiedades estructurales del sistema que la alberga de las cuales depende su comportamiento, y que además el sistema está capacitado para detectar mediante pruebas mecánicas directas. Ambos requisitos están evidentemente ligados

[360] En el marco de concepciones rotundamente externistas de la semántica de lo mental –como la de Millikan (1984)– cobra relieve, no obstante, la idea de que las propiedades semánticas de los estados psicológicos pudiran ser de naturaleza inherentemente histórica. El propio Cummins (1989) ha señalado que la teleosemántica de Millikan se aparta así de la concepción ahistórica de la representación mental que prima en los modelos cognitivos. No en vano, Heil (1989b: 355-356, *infra*) ha recurrido precisamente al tinte histórico que él advierte en la noción de representación mental para justificar que el externismo no es una amenaza para las credenciales de dicha noción en cuanto a la determinación de la conducta ni, consiguientemente, en cuanto a la explicación psicológica –quizá al contrario. Ahora bien, en el contexto en el que Dennett ha planteado la cuestión de la eficacia causal de la semántica, que es el de una ilustración del vocabulario por medio del cual una psicología estrictamente solipsista podría describir estímulos y conductas en términos proximales, es sin duda una concepción ahistórica de la representación la que se presupone.

[361] Desentrañar la analogía puede resultar esclarecedor. Tomando prestadas herramientas de los estudios clásicos de Ogden y Richards (1923), Richards (1936) o Black (1962) sobre la metáfora, cabe replicar que, aun cuando lo que se afirma resulte verosímil en el ámbito del vehículo –las monedas–, sólo contaríamos con razones de peso para pensar que lo sea también en el ámbito del tenor –las creencias, los deseos–, si en el ámbito del fundamento encontráramos una articulación relevante: el hecho de que una creencia sea acerca de algo debe ser, entonces, como el de que una moneda haya estado sobre una mesa, una propiedad histórica. Ésa es, precisamente, la observación de Cummins (1989: 84, *supra*) en la que se basa su argumento de que las representaciones internas postuladas por la teoría computacional de la mente no pueden cabalmente identificarse con las creencias o los deseos.

entre sí: si la conducta del sistema depende de una determinada propiedad, es que el sistema es capaz de detectarla directamente; si solo puede detectarla a través de otra propiedad interpuesta, pronto cobraría fuerza la tentación de concluir –en el mismo espíritu que alienta el desahucio de la semántica del ámbito de la explicación psicológica– que es en realidad esta propiedad interpuesta la que determina el funcionamiento del sistema.

Comoquiera que lo que buscamos son diferencias entre los estados internos de varios sistemas –entre sus representaciones, si hemos de conceder que posean una naturaleza representacional aun estando en discusión que está intervenga activamente en la determinación de la conducta del sistema– de tal naturaleza que, además de ser capaces de ocasionar una divergencia conductual (en la medida, se sobreentiende, que el propio sistema las detecte), no nos sea dado considerar de orden sintáctico –y, dicho sea de paso, que no sean redundantes, que la explicación que nos proporcionen no sea asequible también bajo un planteamiento estrictamente sintáctico–, es de lo más sensato vaticinar que no vamos a encontrar tal cosa, toda vez que los criterios que hemos empleado para acotarla definen una imposibilidad. Puesto que se ha estipulado que propiedades sintácticas de las representaciones mentales son aquellas que resultan relevantes desde el punto de vista de sus conexiones con otros estados internos, incluido aquellos de naturaleza perceptiva o motora, de cara al control de la conducta, se sigue que ninguna propiedad puede ser relevante para la explicación de la conducta sin ser sintáctica. En pocas palabras: las diferencias semánticas que buscamos habrían de desempeñar un papel relevante en la explicación de la conducta, o cual las convertiría en sintácticas, y, a la par, no ser de naturaleza sintáctica ni reemplazables en la explicación por otras que lo fueran. Estamos cazando una quimera: es como si fatigáramos los caminos en busca de una ciudad que se encontrara al norte del paralelo 20°N y al sur del paralelo 10°N –no puede haber tales ciudades, porque hemos demarcado para ellas territorios inexistentes.

Además de la capacidad de determinar la conducta del sistema –"su funcionamiento", dice vagamente Dennett–, la otra condición que definiría el carácter sintáctico de una propiedad es que ha de resultar detectable mediante pruebas mecánicas directas. Es de suponer que esto excluye, por ejemplo, a la referencia o el valor de verdad de la representación mental de una creencia, pues, de nuevo, sólo aquellas propiedades que cuenten con una materialidad local podrían ser objeto de esas pruebas mecánicas directas. La cuestión de la ineficacia causal de la semántica vuelve así a orbitar, implícitamente, sobre la del carácter local de la causalidad, lo que nos remite a observaciones ya trabajadas sobre la diferencia entre causalidad e individuación, o entre el alcance de casos o de tipos que puede dársele a los enunciados sobre relaciones causales: a la luz de esas observaciones, la idea según la cual sólo propiedades físicas locales pueden determinar la conducta del sistema, que como se ha apuntado une las dos condiciones fijadas por Dennett para delimitar su noción de lo sintáctico, resulta tan arbitraria como la de que todo lo causalmente eficaz es sintáctico por definición.

Dando por sentada, entonces, la cuestión de que una propiedad sintáctica deba ser de tal naturaleza que el sistema pueda detectarla mediante pruebas mecánicas directas, lo que tenemos, un poco más cuidadosamente, es que:

Una propiedad física compleja del sistema nervioso es una propiedad sintáctica de una representación mental si y sólo si (*i*) su instanciación genera una variación de las disposiciones conductuales del organismo, sea directa o mediada por la variación de las propiedades sintácticas de otras representaciones mentales.

Pero también que:

Una propiedad física compleja del sistema nervioso es una propiedad sintáctica de una representación mental si y sólo si (*ii*) no es una propiedad semántica, es decir, si los efectos que genere su instanciación son independientes de los hechos a los que se refiera[362].

Y, por otra parte, que:

Una propiedad física compleja del sistema nervioso es una propiedad semántica de una representación mental causalmente eficaz si y sólo si:

(*i'*) su instanciación genera una variación de las disposiciones conductuales del organismo, sea directa o mediada por la variación de las propiedades sintácticas de otras representaciones mentales, y

(*ii'*) las variaciones que genere su instanciación dependen de los hechos a los que se refiera[363].

Pero, como es obvio, cumplir la condición (*i'*), que es idéntica a (*i*), convierte a la propiedad física candidata al rango de propiedad semántica causalmente eficaz en una propiedad sintáctica, mientras que cumplir (*ii'*), que es la negación de (*ii*), la convierte en una propiedad no sintáctica. Las condiciones que hemos establecido para conferir a una propiedad física del sistema nervioso bula de eficacia causal en calidad de propiedad semántica son condiciones contradictorias. Por supuesto, igual que no hace falta ninguna expedición para descubrir que no hay ciudades en latitudes superiores a 20°N e inferiores a 10°N, tampoco hace falta ninguna investigación empírica para saber que, bajo la concepción de la sintaxis y la semántica que venimos describiendo, no hay propiedades semánticas irremplazables en la explicación psicológica. Ninguna de las dos –por lo demás– es una conclusión particularmente interesante[364].

En un lugar muy cercano a éste desemboca también el análisis de las perspectivas de dotar de autonomía explicativa a la semántica de los estados

[362] Donde la relación nombrada por "referirse a" queda, por supuesto, pendiente de elucidar.

[363] Ídem.

[364] *Cf.* Devitt (1989, *supra*) sobre la premisa ilegítima, en el argumento que conduce a la irrelevancia de la semántica, de que toda propiedad relevante de una representación que no sea veritativo-funcional es *ipso facto* de naturaleza sintáctica.

mentales aparejado por Cummins (1989), cuyas líneas maestras se ha expuesto ya. Que el cognitivismo tienda a aceptar que existen generalizaciones en las que se hallan involucrados estados internos identificados en virtud de su semántica, y al mismo tiempo a arrinconar a esa semántica al deslucido papel de trasunto inerte de la sintaxis –a concederle sólo "[…] whatever 'reality' or membership in the natural order goes with having a causal role" (Cummins 1989: 134-135)– no debería suponer –como bien señala Cummins– ninguna sorpresa. Al contrario: más bien se trata de un resultado previsible toda vez que "[…] la estrategia central" del cognitivismo

> […] is to get the formal structures –the representations– to march in step with the things represented. If the symbols track the meanings, the meanings are bound to track the symbols. But the symbols track the meanings because of the causal structure of the device, so the meanings are bound to track the causal relations among the symbols. That content can be used to track the causal transactions of a system will follow trivially from the assumption that the system instantiates a function having those contents as arguments and values. (Cummins 1989: 135)

Esto no son, a fin de cuentas, más que obviedades, y sin embargo –añade Cummins *ibidem*– "[…] the feeling remains that it is the syntax (or the electronics) that does the causing, not the content." Conviene preguntarse, entonces, a qué se debe esa recalcitrante impresión de que la semántica queda relegada. La respuesta de Cummins es que dicha impresión proviene de no haber entendido correctamente los términos en que se produce la articulación de semántica y sintaxis –la formalización. Bajo esos términos, sucede que "[…] el mismo hecho que autoriza a afirmar" que un estado interno tiene un contenido semántico determinado "[…] garantiza" que cualquier transformación semántica de ese estado interno –o, si se prefiere, cualquier efecto que en virtud de sus propiedades semánticas pudiera tener– vendrá reflejada en una transformación sintáctica –en un efecto que se produce en virtud de sus propiedades sintácticas–, y viceversa. Como sucintamente anota Cummins (1989: 135), "[…t]o the extent that this fails, representation fails as well, and you don't have a content to start with".

En efecto, la difícil convivencia entre la rotunda afirmación de que "[…] only symbols have syntax" (Fodor 1985: 27, *supra*) y la no menos tajante de que "[…] syntax reduces to shape" (Fodor 1985: 26, *supra*) nos ha alertado ya acerca de los deslizamientos que parecen operar en la base de la condición de formalidad. Lo que dadas esas dos premisas evita la conclusión indeseada de que cualquier cosa dotada, en algún sentido, de una forma resulte ser un símbolo es, como vimos, que algo sólo posee una sintaxis en tanto que es un símbolo, y sólo es un símbolo en tanto que abriga una semántica. Así que una vez que –supongamos– la semántica de los estados mentales ha quedado limpiamente enhebrada en una sintaxis ya no podemos deshacernos de ella sin asolar también la sintaxis que hemos instaurado: asemejando semántica y sintaxis a dos tableros de un puente –concluye Cummins (1989: 135)–, "[…] you can't 'knock off' the upper span. Its presence is entailed by the structure realized in the lower span". Dicho de otra manera: una propiedad de un símbolo es

sintáctica en tanto no se refiere a nada ajeno al sistema simbólico, pero algo es un símbolo en tanto que se refiere a algo ajeno al sistema simbólico; luego algo llega a tener propiedades sintácticas, en primera instancia, en virtud de su semántica. O, como quedó anticipado *supra*, sólo la presencia de propiedades semánticas habilita la atribución de propiedades sintácticas[365]. Es, acaso, lo que pretende subrayar Martínez-Freire (2001: 91-92) al insistir en que "[…] en el concepto de sistema de símbolos los aspectos semánticos y de interpretación están explícitos" y en que "[…] el sistema de símbolos se concibe como un mecanismo típicamente representacional" para concluir, en suma, que "[…] la distinción entre nivel sintáctico y nivel semántico es forzada" y que "[…] no existe una distinción tajante entre sintaxis y semántica" (Martínez-Freire 2001: 102, *cf*. también Martínez-Freire 1995: 117-118).

Si queremos abrir horizontes un poco menos asfixiantes, seguramente sea preciso que depuremos el criterio para clasificar una propiedad física compleja como propiedad sintáctica o semántica de una representación mental, purgándolo de la condición de relevancia conductual. Éste es –obviamente– el expediente análogo al que empleamos cuando aplicamos la distinción tradicional entre sintaxis y semántica al lenguaje natural, que no acarrea implicaciones acerca de la eficacia causal de las propiedades que la propia distinción deslinda. El curso de la argumentación, en suma, nos devuelve a una pregunta elemental, casi escolar: ¿qué convierte a una característica de un símbolo en una propiedad sintáctica, y qué en una semántica? Partimos, además, de una constatación no menos elemental: sólo cuando ha lugar una interpretación semántica –entiéndase, claro: una interpretación semántica que no sea trivial, que venga arraigada en una explicación de la conducta del organismo o el autómata en su medio– procede considerar sintácticas a algunas de las propiedades físicas de un sistema. Es más: bien cabe aventurar que lo que permite abstraer de entre las propiedades físicas de un símbolo aquellas que constituyen propiedades sintácticas son las propiedades semánticas del símbolo, puesto que no son propiedades sintácticas –ya lo sabemos– todas las propiedades físicas, ni es razonable

[365] Una de las enmiendas a la totalidad que Searle (1990c, 1990d) presenta contra el cognitivismo es la de que cualquier sistema físico que decidamos interpretar como tal –una pared, sin ir más lejos– computa cualquier función que decidamos interpretar que computa, *ergo* la tesis de que los procesos psicológicos son procesos computacionales, o de que el cerebro es una computadora, es trivial. Haciendo uso tácitamente de la misma distinción entre concatenaciones singulares y regularidades que abarcan también escenarios contrafácticos, o entre descripción y explicación, que venimos forjando, Block (1995b) se apresta a rebatir la tesis de Searle con un argumento que entraña también la constatación de que una sintaxis sólo se da en relación a una semántica. El error del argumento de Searle –insistiría Block– es que sencillamente no es cierto que no cualquier entramado arbitrario de propiedades o eventos físicos sea susceptible de preservar el robusto isomorfismo con una interpretación semántica dada que precisamos para poderlo interpretar como una sintaxis –como una sintaxis, para entendernos, con la precisa envergadura contrafáctica–:

> […] the isomorphism that makes a syntactic engine drive a semantic engine […] has to include not just a particular computation that the machine *does perform*, but all the computations that the machine *could have* performed. (Block 1995b: 399)

que lo sean sólo aquellas que están involucradas en la determinación de la conducta o la vida mental del sujeto que alberga dicho símbolo en el marco de sus creencias o sus deseos, sino más bien sólo aquellas –diríamos entonces– que mantienen una determinada relación de correspondencia con una interpretación semántica dada.

Una mirada sobre esta cuestión que comparte varios vértices con la que venimos desentrañando puede hallarse en una pequeña controversia entre McDowell (1994b) y Bermúdez (1995) a cuenta de la atribución de contenido intencional a los estados internos postulados por la psicología cognitiva. En el transcurso de su argumentación, Bermúdez (1995: 364) apunta a la existencia de:

> […] an important equivocation in the concept of syntactic properties. […] Syntactic properties can be understood in purely physical terms, the shape of the letters in a word, for example. Alternatively, we can understand syntactic properties as relational and functional, like the syntactic properties of natural language, as having to do with how words are connected with each other.

Ésta es –qué duda cabe– la misma anfibología que acaba de quedar perfilada, con algunos matices: hemos rastreado siquiera a vista de pájaro el origen de una y otra concepción de lo sintáctico en distintos cauces de la lógica y la lingüística, hemos visto que las propiedades sintácticas en el sentido más rigurosamente formalista no atañen en realidad a rasgos puramente físicos, sino a abstracciones sobre esos rasgos que tienen que ver con su papel en cierta economía de causas y efectos, hemos cuestionado que los criterios sobre los que se funda esa abstracción establezcan la irrelevancia explicativa de lo semántico como no sea trivialmente, por *fiat*.

Sin embargo, pese a la tenacidad con que el cognitivismo ha hecho oir su tesis de que sólo las propiedades sintácticas de los estados internos son relevantes en términos causales, y de que dicha relevancia se explica por el hecho de que estas propiedades sintácticas constituyen abstracciones sobre las propiedades físicas de dichos estados, Bermúdez se muestra convencido de que "[…s]yntactic properties in the second rather than the first sense are relevant […] [to the project of developing a syntactic theory of the brain]". La razón –alega– es que son precisamente los lazos causales entre estados cerebrales –sus roles funcionales, "[…] how they are causally connected up to each other and to sensory input in a way that explains behavioural output" (Bermúdez 1995: 364)– lo que concierne al cognitivista: el paralelismo entre conexiones entre palabras y conexiones entre estados nerviosos queda implícito. Pero el argumento sucumbiría de detenerse aquí, pues obvia que también en términos estrictamente formales la sintaxis tiene que ver con conexiones entre entidades –entre símbolos, si se quiere–, por mucho que se esmere en caracterizarlas sin referencia a otra cosa que algunas de sus propiedades físicas –intuitivamente: su forma. Ése es precisamente uno de los matices recién anotados: que las propiedades sintácticas a las que se refiere la tan traída y llevada condición de formalidad son abstracciones sobre propiedades físicas en virtud de consideraciones causales. Una analogía entre la noción cognitivista y la noción lingüística tradicional de sintaxis fundada sólo sobre la idea de que ambas incumben a las relaciones entre sus objetos –estados

internos, palabras– sería, así pues, demasiado débil como para permitirnos descartar que sea la sintaxis de los inventores de autómatas, y no la de los lingüistas, la relevante para los intereses del psicólogo.

Ahora bien, si consiguiéramos hacer arraigar tal conclusión –que la distinción entre sintaxis y semántica que es decisiva de cara al estudio de la mente es la que media entre la sintaxis y la semántica del lenguaje natural, no la que fuerza a la sintaxis a ceñirse a propiedades físicas de los símbolos–, eso nos dejaría el paso abierto para deshacernos de la tesis de la dispensabilidad de la semántica, ya que "[…i]t is only if syntactic properties are understood in the first sense that a clean break between syntax and semantics seems appropriate" (Bermúdez 1995: 364): si ni siquiera pudiéramos trazar con claridad la linde entre lo sintáctico y lo semántico, difícilmente íbamos a poder pretender que esto se calcara nítidamente sobre aquello al punto de volverse superfluo. Y es exactamente eso –entiende Bermúdez– lo que sucede con la sintaxis y la semántica del lenguaje natural: por ejemplo, que "[…] the syntax of a natural language involves grammatical categories […] [and] one cannot decide what category a given word falls into without speculating about its semantics" (Bermúdez 1995: 364) –un razonamiento muy semejante al que venimos ensayando, aunque enaborlado sin reparar en la existencia de programas reduccionistas o eliminacionistas respecto de la semántica en el seno de la lingüística contemporánea.

Es significativo que el procedimiento por el que el propio Bermúdez logra, más allá del laxo paralelismo entre el carácter relacional de la noción de sintaxis que nos ha legado la lingüística clásica y el de la noción de estados internos propia del funcionalismo, afianzar la tesis de que en psicología como en lingüística las propiedades sintácticas no pueden fijarse íntegramente sin el concurso de la semántica parta precisamente del matiz que apunta al carácter abstracto de las propiedades sintácticas estrictamente formales en relación con las propiedades físicas. En efecto, la pregunta que surge de inmediato al consignar que la psicología cognitiva trata de dar cuenta de relaciones entre estados internos, como la sintaxis entre palabras, es si la identificación de dichos estados internos es factible sin involucrar a sus propiedades semánticas, a diferencia de lo que –según Bermúdez da por hecho– ocurre en lingüística[366]. Aunque la respuesta ha de resultar familiar –lo mismo, en esencia, se ha argumentado *supra*–, conviene citarla *in extenso*:

[366] Es de rigor anotar que Bermúdez (1994: 364) plantea la cuestión en términos de interpretación radical: un intérprete que no conociera ninguna de las propiedades semánticas de las expresiones de una lengua no podría fijar –al menos no del todo, se entiende– sus propiedades sintácticas si no hubiera fijado primero algunas de las semánticas. Aunque éste no es, claramente, el sentido de la noción de interpretación radical inaugurado por Davidson (1973c) en la estela de Quine (1960), la analogía reviste de un aire de cosa probada al supuesto de que la sintaxis de un lenguaje natural no puede llegar a abarcar en su seno toda distinción semántica posible en ese lenguaje –cuando es, como se ha dicho ya, una asunción que ciertas concepciones de la relación entre sintaxis y semántica, basadas como la de Uriagereka (2008) en la tesis de colinearidad, ponen en entredicho.

> […] any such interpretation requires distinguishing inputs with no genuine causal role to play (because they are just noise) from those that have one. Some causal connections are functionally relevant, while others are not, and any satisfactory theory will have to discriminate between these. Second, it is important to identify causal properties that derive from the physical instantiation / realization of a particular state, rather than from its functional role. The problem for the notion of conceiving the mind as a purely syntactic engine is that, although explanation of how the brain works is causal explanation, we need to distinguish relevant causal relations from irrelevant ones. It is clear that syntactic properties as narrowly construed (*i.e.*, in physical terms) will be of no use here. There is no reason to think that there will be any brute physical characteristics demarcating causal properties that are functionally irrelevant. But syntactic properties in the broad sense will be of no use either, because we need to discriminate relevant from irrelevant causal properties in order to fix the broad syntactic properties. So, if we are to discriminate just the functionally relevant causal properties, it seems that we shall have to advert to considerations of function that involve thinking about what can only be described as the semantic issues of how an animal needs to represent its environment. (Bermúdez 1995: 364-365)

Constatar la imposibilidad de fijar criterios para decidir qué propiedades físicas de los estados internos de un sistema cualificamos como sintácticas sin aludir, abierta o veladamente, a las propiedades semánticas de dichos estados conduce en suma –así piensa Bermúdez (1995: 365)– a la convicción fundada de que "[…] a purely syntactic theory of the mind will not be forthcoming". Aunque el propio Bermúdez no parece reparar en ello, el argumento guarda afinidades seguramente más que anecdóticas con las objeciones de Chomsky (1959) al proyecto de explicación del aprendizaje de la lengua en términos de condicionamiento operante esbozado por Skinner (1957), sobre las cuales más de una vez ha orbitado ya el análisis. El papel del conductista, en este trasunto de aquella disputa fundacional, lo desempeña Stephen Stich (1983) y su reinterpretación en tono eliminacionista de la primacía de la sintaxis en la teorización cognitiva[367]:

> Syntactic theories of the mind (like that developed by Stich, for example) owe much of their appeal and plausibility to the parallel between the explanation of a language and the explanation of […] psychological states. […] In the case of natural language we can hive off syntax from semantics, and so it seems a natural thought that a similar, non-intentional approach is possible in explaining, for example, early visual processing. What this neglects, however, is that in the case of natural language we can only hive off syntax from semantics because we have an understanding of the language that includes both its syntactic and semantic aspects. (Bermúdez 1995: 363)

De la misma forma, entonces, que Skinner –según la célebre reseña de Chomsky– introducía veladamente nociones mentalistas de sentido común en análisis de pretendida pureza operacional, Stich –sería la sosegada denuncia de Bermúdez– habría estado mercadeando bajo cuerda nociones semánticas al construir modelos

[367] Reinterpretación en la que el propio Stich ya comenzaba a hallar numerosas grietas (*cf*. Stich 1992, *supra*) y que abandonaría sin tapujos muy poco después: *cf*. Stich (1996).

supuestamente sintácticos de la mente. Así que a fin de cuentas, igual que ocurrió con el conductismo radical, "[…] there are good reasons for thinking that […] a purely syntactic account [of the mind] is a pipe-dream […]" (Bérmudez 1995: 363)[368]. Bien al contrario, si los cognitivistas engarzan en sus teorías psicológicas conceptos intencionales, asumiendo –si se quiere ver así– el *onus probandi* a que ello les obliga, es, mal que le pese a Stich, a Dennett o a Kim, "[…] out of necessity rather than convenience" (Bermúdez 1995: 363).

Esta trabajosa defensa del carácter irrenunciablemente semántico del aparato teórico de la psicología cognitiva sirve a Bermúdez para cuestionar la tesis de McDowell (1994b) según la cual no hay intencionalidad genuina en los constructos explicativos del cognitivismo –en los estados psicológicos subpersonales, dice McDowell siguiendo a Dennett (1969)[369]–, que, a diferencia de lo que ocurre con los estados mentales de la psicología cotidiana –creencias, deseos: estados psicológicos personales, conscientes– carecen de significado *para* el propio sistema y se rigen de manera estrictamente sintáctica. El primer gesto de ese cuestionamiento lo hemos visto ya: Bérmudez no admite que la distinción entre sintaxis y semántica pueda trazarse de tal modo que la semántica quede fuera de nuestros expedientes de identificación de los estados internos que determinan la conducta de un organismo o, llegado el caso, de un autómata.

Hay, además, un segundo reparo que Bermúdez (1995) opone a las tesis de McDowell (1994b), y que tiene que ver con la relación que, de cara a la explicación de la conducta, guardaría esa semántica inconcusa de las creencias y los deseos con la árida sintaxis de los cómputos internos, a la que sólo metafóricamente cabe –asegura McDowell– atribuirle intencionalidad. La posición de McDowell es que esa maquinaria sintáctica crea las condiciones que nos permiten albergar creencias, deseos, esperanzas o temores, y relacionarnos así, intencionalmente, con nuestro entorno. La sintaxis de los cómputos internos no *constituye* entonces la semántica de nuestros estados psicológicos –que lo hiciera sería difícil de sostener toda vez que asumamos la irreductibilidad de la semántica a la sintaxis, y en eso, como hemos visto, ni Bermúdez ni McDowell se muestran dispuestos a transigir–, sino que nos *capacita* para albergarla: "[…] unmysteriously" –dice McDowell (1994b: 200)– "the syntactic virtuosity of our brains enables us to relate to the environment in the direct way that is constitutive of our being the semantic engines that we are". Pues bien, el sentido que Bermúdez concede en este contexto a la idea de condiciones de posibilidad esgrimida por McDowell radica en la idea de que la maquinaria sintáctica postulada por el cognitivismo se presupone en la explicación de la

[368] Etimológicamente, la fantasia de un fumador de opio: es obvio que no es en sentido literal como debe entenderse el giro elegido por Bermúdez.

[369] En realidad, el trabajo de Bermúdez (1995) es una réplica a las críticas de McDowell (1994b) al intento de Dennett (1978a) de construir "[…] una teoría cognitiva de la consciencia", que se basan en la idea de que el propio Dennett desoye las distinciones que había trazado años atrás, en Dennett (1969). No es preciso, a los efectos que ahora mismo nos ocupan, desgranar las tesis de Dennett (1978a).

conducta –como podría decirse también de la gravitación (Bermúdez 1995: 366), o de las constantes físicas: precisamente por razón de su carácter invariante–, la cual se despliega verdaderamente en el ámbito de los estados psicológicos personales. Pero no hay motivos fundados –juzga Bermúdez (1995: 366)– para admitir tal invariancia en lo cognitivo, y hacerlo nos dejaría inermes ante la necesidad de explicar las conductas, a menudo tan intencionales y conscientes como cualquier otra, que surgen cuando el funcionamiento de esa presunta maquinaria sintáctica se desarrolla de forma anormal.

Las dificultades que plantea la propuesta de McDowell, sin embargo, parecen mayores de lo que hace ver el énfasis de Bermúdez en el caso de la conducta anormal. Aislar enteramente el territorio de la explicación psicológica cotidiana del labrado por la teorización cognitiva, depositando en éste nada más que el cometido de dar cuenta del hecho de que podamos atravesar las distintas clases de estados mentales o ejercer las distintas capacidades psicológicas que espontáneamente nos atribuimos, es una operación vacua en tanto no se especifique con qué grado de detalle debe tener lugar. Puede entenderse –esto es– que lo que compete a la teoría psicológica no es sino atestiguar que ciertos mecanismos por ella descritos posibilitan que se den los fenómenos que genéricamente consideramos mentales, o puede entenderse que lo que se le exige es, más bien, que un estado mental de tal o cual clase tienda a producirse, o a derivar en tal o cual conducta, cuando coinciden, digamos, un complejo estimular de estas o aquellas características y un estado mental previo de otra clase determinada –ya se otorgue o no a los estados mentales a los que se apela el reconocimiento de la psicología ordinaria, ya contemos o no con razones para adjudicarle al sujeto consciencia de ellos, o a ellos un contenido intencional definido.

Si lo que se requiere de la teorización cognitiva es sólo que nos tranquilice respecto a que la varia y delicada vida mental que nos atribuimos no es una quimera, sino algo que la maquinaria nerviosa efectivamente hace posible, podremos tal vez articular un modo de esclarecer esas condiciones de posibilidad que no entorpezca la autonomía de las explicaciones planteadas bajo los esquemas de la psicología cotidiana, pero será a costa de que para los pormenores de esas explicaciones subsista la pregunta por las condiciones que pudieran rendir cuenta de que seamos capaces de mostrar esa faceta específica de nuestra relación con el entorno. Si, en cambio, reclamamos de la ciencia de la mente que nos permita entender cómo es posible que cada pensamiento o cada motivación de una índole determinada se produzca en las condiciones precisas en que se produce, y no en otras –o también, como apunta Bermúdez (1995: 366, *supra*), que existan estas o aquellas quiebras de la normalidad, en las que ciertas ideas, deseos o conductas, pero no otras, se vean malogradas–, entonces el número de preguntas sin respuesta irá poco a poco menguando, pero lo hará al mismo ritmo que el margen de autonomía de unas explicaciones psicológicas de naturaleza personal a las que hemos desprovisto de la posibilidad de entretejerse con la teorización sobre estados internos subpersonales, y que por tanto sólo podrán cederles el paso. Lo más sensato es pensar, seguramente, que lo que debamos requerir de una teoría madura sobre lo mental no es ni un

recuento exhaustivo de la genealogía de cada uno de los estados psicológicos que la psicología ordinaria nos permita distinguir –no es prudente, como bien ha señalado Putnam (1997: 35, *supra*), aspirar a la formulación de "[…] a set of laws that distinguish, say, the state of being jealous of Desdemona's fancied regard for Cassio from every other actual or possible propositional attitude"–, ni tampoco un mero salvoconducto genérico para explicar nuestra propia conducta y la de nuestros semejantes –humanos o no– en términos de estados mentales conscientes y sus contenidos, sin atender al modo en que la misma arquitectura interna que posibilita, en general, que abriguemos tales estados mentales está involucrada en que, en particular, unos u otros intervengan en la determinación de la conducta que tratamos de entender. Ahora bien: sea cual sea el grado de detalle que proceda exigir a una teoría psicológica –cuestión en torno a la que no es difícil prever tintes pragmáticos–, es de esperar que se reitere una y otra vez este mismo equilibrio entre la autonomía de una comprensión de la conducta en términos personales y la carestía de explicaciones detalladas de las condiciones que posibilitan unos comportamientos y no otros. Dicho con otras palabras: delinear dos niveles de explicación –el de la psicología cotidiana, que apelaría a actitudes de la persona, como las creencias o los deseos, y el de las ciencias cognitivas, que postularían estados mentales subpersonales a los que se atribuirían algunos de los rasgos de aquellas actitudes, aunque no siempre todos– y aislarlos entre sí a la manera que McDowell bosqueja –a saber, concediendo la condición de formalidad según la cual estos últimos se definen en términos estrictamente sintácticos, y reservando el privilegio de lo semántico para los conceptos de la psicología cotidiana– nos aboca a debatirnos entre, de un lado, una versión del "problema de la trivialización" que según Fodor y Pylyshyn (1981) aqueja a todo ensayo de realismo directo (*cf*. Pylyshyn 1984: 182-183, *supra*), ya sea de estirpe gestáltica, conductista o gibsoniana[370], y, del otro, la sumisión última de toda explicación de índole personal, intencional, a una subpersonal, cuya intencionalidad es, de acuerdo con los términos en los que las hemos aislado entre sí, meramente parasitaria. O la apelación a la intencionalidad –a la semántica–, en definitiva, sufre de indigencia o de dispensabilidad: en ninguno de los lugares a los que pudiera llevarnos el camino que McDowell señala se alumbra el engranaje entre causas y razones que hemos andado buscando.

Un ensayo de restitución

Las vacilaciones en torno a la eficacia causal de la semántica no serían a ojos de Devitt (1989) sino rastros, digamos, girondinos de una corriente revisionista más amplia –"the Revisionist Line"[371]– cuya facción jacobina sería, como piensan también Fodor (1989, *supra*) o Toribio (1991, *supra*), el eliminacionismo, y que convive en las ciencias cognitivas con la reivindicación de la psicología natural –"the Folk Line". A

[370] No quiere decirse con esto, por supuesto, que McDowell haya por fuerza de convenir en ninguna de estas ideas de la percepción. Sobre el caso de J.J. Gibson en particular, *cf*. McDowell (1994a: 47)

[371] La noción de *revisionismo* acerca de la naturaleza de lo mental, tal como se aplica en este contexto, puede verse ya plasmada en Burge (1986, *supra*).

su entender, el revisionismo estaría estrechamente ligado a la búsqueda de un territorio autónomo para la explicación psicológica propiciada por el funcionalismo y la teoría computacional de la mente, y en particular a la densa constelación conceptual en que concurren solipsismo metodológico, semánticas funcionales, internismo y contenido restringido:

> This [revisionist] line frequently holds that cognitive psychology must explain the interaction of mental states with each other and the world by laws that advert only to formal or syntactic properties, not to truth-conditional ones. (Devitt 1989: 369)

La reivindicación del esquema explicativo de las actitudes proposicionales, por el contrario, vendría inspirada por las tesis externistas respecto a la naturaleza de la explicación psicológica y la semántica de los estados mentales, basadas en la noción de contenido amplio:

> More particularly, the inspiration is the folk view that people have thoughts with rich representational and semantic properties; in particular, people have mental states with truth-conditional content. Cognitive psychology must explain the interaction of thoughts with each other and the world by laws that advert to these semantic properties […]. (Devitt 1989: 369)

Perfilar este mapa lleva a Devitt a un análisis de las nociones de propiedades sintácticas y propiedades semánticas empleadas en la contraposición entre las tesis internistas y las externistas; este análisis le servirá, a su vez, para construir una interpretación del funcionalismo que, a su juicio, resulta inmune a las tentaciones epifenomenistas o eliminacionistas, sin vulnerar los márgenes del internismo[372].

Dos son para Devitt los principales argumentos que respaldan la tesis revisionista de que las leyes de los procesos mentales deben apelar únicamente a propiedades sintácticas de las representaciones que en ellas se postulen: uno parte de la analogía computacional, otro del solipsismo metodológico y la búsqueda de una psicología autónoma. Ambos son, a su juicio, fácilmente revocables. La analogía computacional no establece que *todas* las propiedades de las representaciones mentales que intervengan en las leyes psicológicas deban ser sintácticas porque el alcance de la analogía se restringe a los procesos que conducen de un estado mental a otro –ambos simbólicos. Escaparían de la analogía en cambio –piensa Devitt– los procesos que conducen de estímulos a estados mentales o de estados mentales a conductas, donde tanto estímulos como conductas son en gran medida de naturaleza

[372] Que la matizadisima abdicación del solipsismo metodológico por parte de Fodor (1990) y su tránsito hacia el externismo fueran acompañados por un persistente realismo acerca de las actitudes proposicionales resultaría fácil de comprender a la luz de la cartografía trazada por Devitt. Lo mismo sucedería, como señala Devitt (1989: 370), con la acusación que Stich (1983) vertía contra Fodor por tratar de avenirse a dos enfoques de la explicación psicológica tan contradictorios entre sí, al menos a primera vista.

no simbólica, a diferencia del *input* y el *output* de los procesos computacionales. Como una consecuencia de esto ve Devitt el hecho de que:

> Any interpretation a computer's symbols have, *we* give them. However, it is plausible to suppose that a human's symbols have a particular interpretation in virtue of their perceptual causes, whatever we theorists may do or think about them. Furthermore, it is because of those links to sensory input that a symbol has its distinctive role in causing action. (Devitt 1989: 375)

En cuanto al solipsismo metodológico: la posición en que Devitt se atrinchera es que incluso si fuera cierto que el papel explicativo que esperamos de las representaciones postuladas en las leyes psicológicas sólo pueden desempeñarlo propiedades que supervengan en estados internos del organismo, aun así lo único que quedaría establecido es que las propiedades veritativo-condicionales de dichas representaciones son inútiles para el trabajo en cuestión –puesto que, obviamente, establecen lazos referenciales con el entorno que no cumplen ese criterio de superveniencia[373]–, pero en ningún caso se habría apuntalado la conclusión de que sólo las propiedades sintácticas tengan cabida en psicología. La premisa mayor –que toda propiedad relevante de una representación que no sea una propiedad veritativo-funcional es una propiedad sintáctica– está oculta y, a juicio de Devitt, es claramente falsa[374].

Así pues, la noción de contenido restringido se despliega en manos de Devitt como un intento de establecer que determinadas propiedades de las representaciones

---

[373] Todo esto –reiterémoslo– concediendo que el solipsismo metodológico fuera sólido. Pero es que, además, Burge (1986, *supra*) habría demostrado fehacientemente que sus postulados fundamentales son erróneos, pues en ellos se transita confusamente desde una inocua constatación del carácter local de la causación hasta una injustificada tesis que otorga carácter local –interno– también a la individuación. Aunque Devitt acepta el argumento de Burge, considera sin embargo que éste no desarticula por completo el solipsismo metodológico, en tanto que trazar unos determinados límites al tipo de fenómenos que incluiremos en la explicación de nuestro objeto de estudio es una decisión legítima –precisamente, una decisión metodológica–, y los límites que marca el solipsismo metodológico están a juicio de Devitt donde deben estar:

> I think that underlying the argument is the strong conviction that a scientifically appropiate boundary for explaining the behavior of an organism is its skin. It is by stopping at that point that we shall get the appropiate laws. I share this conviction. (Devitt 1989: 388)

Sobre la referencia a la piel como frontera natural entre el organismo y su entorno, *cf*. Skinner (1974: 17), Block (1986: *passim*), Place (1999: 380), *supra*.

[374] Que la expulsión de la semántica del ámbito de la explicación psicológica se ha obrado, en un sentido estrechamente relacionado con esta observación de Devitt, por *fiat* es una idea que se ha examinado *supra*.

El razonamiento de Devitt nos apremia, así pues, a la búsqueda de propiedades de las representaciones postuladas en la explicación psicológica que supervengan en estados internos del organismo, pero que no sean de naturaleza sintáctica: a su entender, la noción de contenido restringido –o de contenido proto-veritativo-condicional, en las palabras del propio Devitt: *cf. infra*– cubriría exactamente la minuta.

postuladas en la explicación psicológica no son de naturaleza sintáctica pese a que supervienen en estados internos del organismo, bloqueando de esa manera la interpretación revisionista –en el extremo, eliminacionista– del carácter solipsista que a su entender debe mantener la semántica de las ciencias cognitivas. Una mirada en profundidad a lo que entraña la noción de contenido restringido –piensa Devitt– revelará que su carga semántica es mucho mayor de lo que pudiera hacer pensar una descuidada identificación con lo sintáctico.

> Narrow meaning (or content) is very rich. Not only does it include all the functional roles that determine the syntactic structure of a sentence, but also the inner functional roles that partly determine the reference of words. […] The latter roles are what is left of wide word meaning when the extra-cranial links are subtracted. The roles constitute narrow word meanings. Those meanings are functions taking external causes of peripheral stimuli as arguments to yield wide (referential) meanings as values. […] *Narrow word meaning is (mostly) not a matter of syntax*. (Devitt 1989: 378)

La importancia de la noción de contenido restringido radica entonces, a ojos de Devitt, en que el carácter veritativo-funcional de la semántica en sentido pleno –*i.e.*, del contenido amplio– encuentra un límpido reflejo en el carácter proto-veritativo-funcional del contenido restringido, que dista de constituir un mero conjunto de propiedades sintácticas[375]. En efecto:

> [...T]ruth-conditional meaning […] is clearly wide, for it partly depends on causal links that are 'outside the skin'. If we abstract from those outside links, we are left with "proto-truth-conditional" meaning. This meaning is narrow in that it is entirely supervenient on the intrinsic inner states of the thinker. It is the inner functional-role *part of* wide meaning. (Devitt 1989: 378)[376]

---

[375] Al afirmar que las propiedades semánticas restringidas no son propiedades sintácticas en la medida en que no atañen sólo a la estructura de la proposición –la oración, la fórmula del lenguaje del pensamiento… –, sino que además contribuyen a fijar la referencia de sus constituyentes, Devitt prefigura, de nuevo, las matizaciones que venimos abordando acerca del peculiar sentido que el concepto de sintaxis ha adquirido en el funcionalismo, habida cuenta de la confluencia en su seno de ideas de lo sintáctico provenientes de la teoría de la computación y la lógica formal, pero también de la lingüística.

[376] De hecho, Devitt opta por mencionar el contenido restringido de un estado mental mediante un artificio formal que decreta la extirpación de sus lazos referenciales y, con ello, de su carácter (plenamente) veritativo-funcional: así, si el contenido amplio de una creencia es *Fa*, su contenido restringido es **Fa*. El propio Devitt (1989: 397) recalca que tanto McGinn (1982) como Loar (1982) parecen ver el factor de rol funcional de sus teorías semánticas desligado del factor veritativo-funcional; también Block (1986: 86, *supra*) considera una confusión pensar que el significado restringido sea una parte del contenido amplio.

Conviene advertir que Fodor, según relata Block (1986: 92), considera la idea de que el contenido restringido es de naturaleza semántica una forma de la falacia de substracción: "[…] assumming that, if you take meaning or content and *substract* its relation to the world and its social aspect, what you have left is something semantic". La objeción, *a fortiori*, dañaría el planteamiento de Devitt si es que daña el de Block, pues lo que Devitt parece sugerir es que la carga semántica del contenido restringido es mayor incluso de lo que Block reconoce. Además, la réplica de Block (1986:

Con estos aparejos, Devitt (1989: 381) dispone la defensa de su noción de contenido restringido en dos frentes: se trata, por una parte, de mostrar que el contenido restringido es *necesario* en la explicación psicológica –para desmantelar así la teoría sintáctica de la mente– y, por otra, de hacer ver que es *suficiente* –haciendo superfluo, por tanto, el proyecto de una psicología externista inspirada en la noción de contenido amplio.

El argumento a favor de la necesidad de la noción de contenido restringido depende crucialmente de las tesis de Devitt sobre las fronteras de la sintaxis. Según su planteamiento, a qué tipo de oración –pongamos por caso– pertenezca *Fa* no es una propiedad sintáctica, aunque sí lo sea su pertenencia al mismo o distinto tipo de expresión que *Fb*. Al abordar un proceso mental tan elemental como la formación de la creencia de que *Fa* a causa de la percepción de determinados estímulos, nos es preciso entender por qué esos estímulos en particular conducen a *Fa*, y no a cualquier otra creencia. Pero si nuestra psicología únicamente admite propiedades sintácticas de *Fa* –propiedades estrictamente formales, como ser un predicado monádico, o ser diferente de *Fb*–, no nos permitirá aislar *Fa* como la creencia causada por esos estímulos: también *Ga* o *Gb*, por dar sólo un par de casos, son predicados monádicos distintos de *Fb*. La noción de contenido restringido resuelve parsimoniosamente el problema, pues nos permite apelar a los lazos de *Fa* con la estimulación proximal responsable de la percepción del estímulo en cuestión. *Mutatis mutandis*, la explicación debería valer para las leyes que describen el vínculo entre creencias y conductas[377]:

> These [laws] must explain the fact that a thought led to a certain behavior and not to others. […] How could the role of […] [a given] belief possibly be explained if we ascribe to it only syntactic properties? Why should that belief lead to this behavior rather than [any other] […]? A thought has a distinctive role in producing behavior […] Syntax alone cannot explain that distinctive role. (Devitt 1989: 382)

Desde aquí, la labor de Devitt se fragua en el intento de obstaculizar una respuesta que Stich (1983) anticipa contra esta tesis de que la teoría sintáctica de la mente no nos facultaría para dar cuenta de las relaciones entre nuestros estados mentales, los estímulos que los originan y las conductas a las que dan lugar. A tal efecto, Devitt

---

92) –que se muestra dispuesto a conceder la cuestión léxica de que el contenido restringido no deba calificarse de "semántico", mientras se le otorgue que es "[…] a distinct feature of language, a characterization of which has something important to contribute to a total theory of meaning"– está, por motivos obvios, fuera del alcance de Devitt.

La comparación de la noción de contenido restringido de Devitt con las distintas reconstrucciones del mismo concepto que se emplean en las semánticas bifactoriales –aquellas que distinguen significado restringido y significado amplio– abre horizontes mucho más extensos de los que cabe otear aquí.

[377] Pero *cf.* Devitt (1989: 389, 393-394, *infra*) sobre el problema de la descripción proximal y distal de la conducta.

nos advierte de que los bosquejos de leyes que Stich propone para salvaguardar el vigor explicativo de la sintaxis son desmedidamente específicos y, sobre todo, de imposible generalización. Las propuestas de Stich a las que Devitt (1989: 383) alude son, por ejemplo:

> [...] For all subjects S, when an elephant comes into his view, S will typically come to have a sequence of symbols, E, in his B[elief]-store.
> […] For all subjects S, if S has […] the sequence of symbols R in his D[esire]-store […], and if S has no stronger incompatible D[esire]-states, then S will raise his arm. (Stich 1983: 178-179)

Las protestas de Devitt (1989: 383) son indudablemente justificadas: estas presuntas leyes pasan por alto la estructura composicional de las secuencias de símbolos que mencionan, y se apartan así del "nivel teórico correcto" para una explicación convincente de la conducta:

> This is indicated by the fact that psychology would require indefinitely many such laws to cope with the indefinitely many possible stimuli for, and behavioral outcomes of, beliefs. […] [They] are at best low-level laws that are applications of [the] higher-level laws […] that we need. (Devitt 1989: 383)

Se diría que Devitt es, sin embargo, indulgente, al no señalar el hecho de que las leyes esbozadas por Stich contienen explícitas menciones de estímulos distales y de acciones, lo que de por sí las vuelve irreconciliables con la clase de explicaciones sintácticas que Stich anhela.

En cualquier caso, lo medular de la crítica de Devitt es que Stich, con los recursos que se ha dado, no puede diferenciar apropiadamente entre *E* –la secuencia de símbolos ligada a la visión de un elefante– y cualquier otra, como *T* –la secuencia de símbolos, digamos, ligada a la visión de un tigre. Aparte de que es muy probable que ambas secuencias tengan idéntica estructura sintáctica, el caso es que aunque sean diferentes tipos de creencias y esa diferencia sea de naturaleza sintáctica, para poder diferenciarlas apropiadamente necesitaríamos, de acuerdo con la doctrina de Devitt, saber a qué tipo en particular corresponde cada una, lo cual no es ya una propiedad sintáctica: para saberlo, claro, nos bastaría con anudar una secuencia de símbolos a la presencia de un estímulo y otra a la del otro, pero eso es semántica, y semántica es lo que Stich aspiraba a vetar.

Con todo, la debilidad del argumento de Devitt se condensa precisamente en su tesis de que la pertenencia de una representación a una u otra categoría psicológica no es una propiedad sintáctica, que es demasiado endeble para la carga argumentativa que deposita en ella. De hecho, Devitt plantea el argumento en términos lingüísticos –la pregunta es, pues, a qué categoría pertenece una oración–, lo cual, como hemos visto ya, genera un intersticio respecto al significado de los estados mentales que no necesariamente es fácil salvar. En suma: que las propiedades semánticas de la *lingua mentis*, a diferencia de lo que acaso ocurra en el

lenguaje natural, se pliegan a sus propiedades sintácticas es la hipótesis que el cognitivismo extrae de la reflexión sobre la naturaleza de los procesos computacionales; poco avanzamos, como se ha visto, por señalar que esa hipótesis no parece, a simple vista, cumplirse en el caso de determinadas oraciones del lenguaje natural. Por otra parte, ni siquiera está claro a qué se refiere Devitt cuando habla de categorías de oraciones: obviamente, si se refiere a categorías formadas por criterios sintácticos, su tesis de que la pertenencia a tales categorías no es una propiedad sintáctica es un mero contrasentido; pero si refiere a categorías formadas por criterios semánticos —como sus ejemplos dan a entender, puesto que "Hay un elefante allí" pertenecería a distinto tipo de oración que "Hay un tigre aquí"– entonces su tesis es una mera tautología.

En cambio, cuando Devitt delimita la analogía computacional apelando al carácter no simbólico de buena parte de los estímulos y conductas que se entrelazan con nuestros estados mentales –a diferencia de los *inputs* y *outputs* de los procesos computacionales, tanto los estímulos como las conductas que conciernen a la explicación psicológica son, decíamos, en gran medida de naturaleza no simbólica (*cf. supra*), es cuando su rechazo a tratar las relaciones entre estímulos y actitudes proposicionales –o entre actitudes proposicionales y conductas– como asuntos sintácticos suena más convincente. Las relaciones sintácticas, al fin y al cabo, se dan entre símbolos, no entre símbolos y otras cosas. Sin embargo, el propio Devitt debilita esta línea de razonamiento al insistir en mantenerse dentro de las lindes del internismo: su noción de contenido restringido nos autoriza a interpretar el vínculo entre la actitud proposicional y el estímulo proximal –o la conducta– como relaciones entre símbolos, y, por tanto, como relaciones que cabe repensar en términos sintácticos. Sólo nos hace falta, para ello, trazar nuestra frontera metodológica en la vertiente interna de transductores o efectores, en lugar de en su vertiente externa –o, como dice Devitt, en la piel–, de modo que los estados físicos de dichos mecanismos cuenten como embrionarios símbolos de la estimulación sensorial o de las respuestas motoras. Y trazar dicha frontera allí donde nos parezca propicio es un derecho que el propio Devitt ha reconocido: "[...i]t is possible to draw boundaries anywhere and to look for explanations of the characteristics and peripheral behavior of the bounded entity or system in terms of what goes on within the boundary" (Devitt 1989: 388). Dicho de otro modo: el esfuerzo de Devitt de arraigar a mitad de camino entre el revisionismo de inspiración internista y el externismo es ineludiblemente engullido por el solipsismo. El terreno es más cenagoso de lo que Devitt creía.

Queda por entender, pues, la renuencia de Devitt a adentrarse en el externismo. El segundo frente en que se desarrolla su argumento es –como se ha dicho– precisamente ese: defender que la noción de contenido restringido es suficiente para los propósitos de la explicación psicológica, y el externismo, por tanto, una prolijidad poco elegante. Dos, y desparejos, son los puntales de esa defensa. De un lado está la convicción de Devitt, que ya se ha mencionado, de que el lugar señalado para poner coto a la explicación psicológica es la piel, pero esto no es más que una convicción. Del otro lado, Devitt invoca a la intuición solipsista de que

el mundo real es irrelevante si lo que nos guía son las representaciones que nos formemos de él –la intuición nuclear del internismo respecto de la explicación de la acción, que hemos examinado de la mano de Guttenplan (1994b: 290, *supra*):

> Cognitive psychology seems to be concerned with mental states that purport to represent […] [a] world which is external to the mind and toward which the organism's behavior is directed. It is of the essence of those mental states –thoughts– that they do purport to represent a world. […] NARROW PSYCHOLOGY does not deny the importance of representation. It simply claims that psychology does not care *which* entities the organism, as a result of its external links to the environment, is actually representing; indeed, it does not care whether there really are such entities. What matters is only how the world seems from the point of view of the organism. The nature of a mental state is indeed 'outward-looking'. It just does not matter to psychology which world, if any, the state actually 'sees'. (Devitt 1989: 389, mayúsculas en el original)

Pese a que su planteamiento concierne de forma tan crucial a la forma en que los hechos mundanos a los que estos se refieren deban quedar caracterizados en la descripción de nuestras creencias y deseos, Devitt posterga hasta las últimas páginas de su trabajo el momento de hacer frente al problema de la especificación, proximal o distal, de estímulos y conductas, que hemos abordado *supra*. Las consecuencias que la respuesta a esa pregunta puede encerrar para su concepción de la explicación psicológica no son, desde luego menores:

> If behavior is described in ways that refer to things outside the skin, as it is by the folk, then it will have to be explained by mental states that refer to such things also. […] The appropriate boundary for psychology will not be the skin. (Devitt 1989: 389)

Ahora bien, Devitt se muestra dispuesto a conceder a Burge –*cf.* Burge (1986: 44-45, *supra*)– que:

> What psychological laws explain is not behavior described as neural impulses, as mere bodily movements, nor as any other brute-physical event. These descriptions are at the wrong level, the level of psychological *implementation*. The level that yields the interesting generalizations of psychology requires that the behavior be treated as an action […]. (Devitt 1989: 393-394)

El propio Devitt (1989: 382), en el transcurso de su controversia con Stich (1983), emplea como ejemplos de la clase de conductas que estarían llamadas a figurar como consecuentes de las leyes psicológicas cosas tan poco afines a una descripción física proximal como *abrir una puerta que se interpone en el camino de un jinete* o *quitar del alcance de una mujer cualquier objeto al que ella pudiera dar un patada*; otro tanto ocurre, desde luego, respecto de la descripción de los estímulos que ocasionaron los estados internos conducentes a tales conductas.

La lección que Devitt extrae de esto es, simplemente, que la ciencia cognitiva no puede ser una ciencia reduccionista a la antigua usanza. Pero aunque atacar a los viejos positivistas lógicos sea a menudo una treta retórica eficaz, Devitt debe ver que

la inviabilidad de la descripción proximal de la conducta genera una brecha en su noción de contenido restringido, en la medida en que los criterios utilizados para describir estímulos y respuestas no respetan las estrictas restricciones que se han impuesto a la descripción de los estados internos con los que dichos estímulos y respuestas se engarzan en complejas cadenas de relaciones funcionales. Una vez que ha cedido a Burge en cuanto a la descripción de estímulo y respuesta, el único camino que a Devitt le queda abierto es, así pues, renunciar también a la descripción proximal de los contenidos de los estados internos, o lo que viene a ser lo mismo, a la noción de contenido restringido[378].

El resto del trabajo de Devitt es un intento de repartir las afinidades de la psicología natural entre la noción de contenido amplio –a la que normalmente se le adjudican– y la de contenido restringido. Que ese elaborado sentido común al que llamamos psicología natural emplee a menudo alguna forma de la noción de contenido restringido es probablemente cierto, y los argumentos de Devitt lo corroboran en buena medida. Pero esto no ayuda a decidir la cuestión de si a la psicología científica debe o no vedársele una semántica plenamente referencial, instaurada sobre la noción de contenido amplio, ni tampoco la de si, como pretende Devitt, cabe articular una semántica restringida que no recaiga del todo sobre la sintaxis.

Naturaleza en la naturaleza

En un trabajo de vastísimo aliento titulado concisamente *Mente y mundo* deja clara McDowell (1994a) su denuncia de que la presunta indigencia de la semántica de creencias y deseos en la explicación de nuestras acciones, cuyos perfiles dibujábamos *supra*, consiste en esencia en la incapacidad de responder a preguntas que de todos modos carecen de sentido:

> Muchos trabajos contemporáneos se proponen, con el espíritu propio del naturalismo, responder (y no exorcizar) preguntas que se pueden encuadrar dentro de la fórmula interrogativa '¿Cómo es posible...?', ya sea acerca del contenido empírico u otros aspectos de lo mental. Los trabajos a que me refiero tratan de proporcionar descripciones perspicuas del modo en que están constituidos materialmente los, digamos, sujetos perceptivos, de tal modo que se nos haga inteligible el que cosas compuestas meramente de materia como ellos puedan poseer el oportuno complejo de capacidades preciso como para que sean capaces de percibir. (McDowell 1994a: 30)

[378] No es raro –así visto– que la noción de contenido restringido de Devitt se acerque a la de McDermott (1986, *supra*). Sin embargo, Devitt menciona su vecindad con McDermott sólo en un punto: que la referencia de un estado mental no es la estimulación proximal, sino la distal. Dicho de otro modo: es la teoría del contenido restringido de los estados mentales la que se refiere a la estimulación proximal, no los propios estados mentales –como no sea, claro, desde el punto de vista de la especificación de su contenido restringido (Devitt 1989: 397).

Pero estas respuestas –arguye McDowell– son inútiles, porque las preguntas a las que se dirigen son el fruto comprensible pero espurio de ciertas "[…] angustias distintivas de la filosofía moderna" (McDowell 1994a: 23). Lo que debe hacerse con tales preguntas no es responderlas sino más bien –ya se ha dicho– *exorcisarlas*: entender por qué nos hemos sentido obligados a responderlas, por qué incluso, como a Fodor (1987: *xiii*, *supra*) o a Dretske (1988: *x*, *supra*), hacerlo nos ha parecido perentorio, o por qué, como a Huxley (1866, 1870, *supra*) o du Boys-Reymond (1872, *supra*), se nos pudo antojar que quedaba fuera de nuestro alcance.

La visión del mundo que subyace al esfuerzo por detallar las condiciones que hacen posible que la semántica de los estados mentales desempeñe un papel en la determinación de la conducta es –apunta McDowell 1994a: 178)– "[...] una filosofía cuyo proyecto es el de empezar a partir del mundo natural, y hacer luego un sitio en él para las mentes y sus contenidos". El resultado de plantearse así la tarea se hace, una vez que dicho planteamiento ha quedado sometido a cierto escrutinio crítico, casi previsible:

> La filosofía moderna ordinaria se enfrenta de un modo bien característico a los dualismos que de ellas se derivan. En primer lugar, asume una posición en una orilla del abismo que trata de franquear, aceptando sin cuestionamiento alguno el modo en que el dualismo al que se enfrenta define esa orilla del abismo. A continuación, construye algo que quede tan cercano como sea posible a la concepción acerca de la otra orilla que figura en los problemas, a partir de materiales que están disponibles de manera no problemática en la orilla sobre la cual ha asumido su posición inicial. Naturalmente, al final ya no parece que siga existiendo un abismo, pero el resultado tiene que parecer por fuerza más o menos revisionista (más revisionista cuanto más urgentes fueran los aparentes problemas originarios, cuanto más firmemente atrincherado se encontrase el modo de pensar que provocó la apariencia de un precipicio infranqueable). (McDowell 1994a: 156-157)

En particular, lo que subyace a la pregunta por la imbricación de la intencionalidad en explicaciones psicológicas de inspiración naturalista, o mecanicista –como a muchas otras de esas angustias filosóficas– es, según el penetrante diagnóstico de McDowell:

> […] una tensión entre dos fuerzas […] que tienen la comprensible tendencia a configurar cada una de ellas a su manera nuestra reflexión en torno al pensamiento empírico –y, por lo tanto, nuestra reflexión en torno al hecho de dirigirse al mundo en general. Una de tales fuerzas es el atractivo que posee un empirismo mínimo, el cual implica que la idea misma de que el pensamiento se dirija hacia el mundo empírico es inteligible sólo en términos de responsabilidad ante el tribunal de la experiencia –concebida ésta, a su vez, como las impresiones procedentes del mundo que inciden sobre los sujetos perceptivos–. La otra fuerza es un modo de pensar que hace que parezca imposible que la experiencia sirva de tribunal. La idea de tribunal, junto con la idea de aquello sobre lo cual el tribunal emite sus veredictos, pertenece a lo que Sellars llama "el espacio lógico de las razones" [*cf. supra*]: un espacio lógico cuya estructura consiste en que algunos de sus elementos, por ejemplo, estén justificados, o sean correctos, gracias a otros. Pero la idea de

> experiencia, al menos si esta se entiende como algo formado por impresiones, pertenece evidentemente al espacio lógico de las conexiones naturales. (McDowell 1994a: 23-34)

Como es de esperar que se hubiera hecho ya diáfano en la discusión de la condición de formalidad estipulada por Fodor (1985) para la explicación psicológica, sobre pensamientos, creencias o juicios pesan así pues dos exigencias contrapuestas: que su naturaleza se agote en la descripción de sus relaciones causales, mecánicas, en su sintaxis –que se integren en el reino de la ley, de lo que consideramos la explicación natural–, y al mismo tiempo que se plieguen a la normatividad que, en tanto que fenómenos intencionales, no podemos dejar de reconocerles, pues sabemos de cierto que "[…u]na creencia o un juicio de que las cosas son del tal o cual modo […] debe ser una actitud o postura que se adopta *correcta* o *incorrectamente* en función de si las cosas son efectivamente de tal o cual modo […]"; es decir, que "[…] el pensamiento […] se hace responsable [*answerable*] ante el mundo –ante cómo son las cosas […]" (McDowell 1994a: 15-16) –es decir, que se ciñan también al espacio lógico de las razones.

Escapar de ese atolladero requiere a juicio de McDowell impugnar los términos en que se nos plantean las preguntas. En particular, debemos rechazar tanto que a la sensibilidad, a las impresiones sensoriales que ubicamos en el espacio lógico de las causas y los efectos, le corresponda unívocamente la pasividad como que al entendimiento, a la formación de pensamientos y juicios en el espacio lógico de las razones, por medio de la cual le sería dado a la sensibilidad participar en ese tribunal de la experiencia al que un empirismo mínimo la convoca, le corresponda unívocamente –con Kant– la espontaneidad. El ejercicio conceptual que vinculamos a la espontaneidad –insiste McDowell a cada paso– impregna también el ámbito de la sensibilidad:

> Es preciso que concibamos esta espontaneidad así extendida como algo sujeto al control desde fuera de nuestro pensamiento, so pena de que, de no hacerlo así, terminemos por representarnos las operaciones de la espontaneidad como una rueda que gira en el vacío sin fricción alguna con el exterior. (McDowell 1994a: 47)

Nuestras capacidades conceptuales, así pues, operan en nuestra experiencia, pero no lo hacen con la espontaneidad que según creemos sin cuestionarlo les es irrenunciable, sino pasivamente; aún así, son esas mismas capacidades conceptuales –no un trasunto suyo apenas reconocible, lo que impediría trabar la fricción del pensamiento con el mundo que tratamos de discernir:

> No seríamos capaces de suponer que las capacidades que entran en juego en la experiencia son capacidades conceptuales si tales capacidades fuesen algo que se manifestase únicamente en la experiencia, únicamente en las operaciones de la receptividad. No se las reconocería como capacidades conceptuales en absoluto de no ser porque podrían ejercerse asimismo durante el pensamiento activo, es decir, de diversas maneras que parecen casar bien con la idea de espontaneidad. [...] De forma bastante

> general, podemos decir que las capacidades que se aprovechan en la experiencia se reconocen como conceptuales únicamente cuando se las considera sobre el trasfondo del hecho de que alguien que las posea ha de ser capaz de reaccionar ante las relaciones racionales que ligan los contenidos de los juicios de experiencia con otros contenidos susceptibles de entrar en los juicios. (McDowell 1994a: 48)

De lo que se trata, en suma, es de que "[…] la experiencia, aun siendo algo pasivo, hace que se manejen capacidades que genuinamente pertenecen a la espontaneidad" (McDowell 1994a: 51), o bien, expresado con mayor solicitud, de que:

> Las experiencias, ciertamente, se constituyen por operación de la receptividad, de forma que pueden satisfacer la necesidad de un control externo sobre nuestra libertad en el pensamiento empírico. Pero las capacidades conceptuales, las capacidades que pertenecen a la espontaneidad, operan ya en las experiencias mismas, no sólo en los juicios basados en ellas: de manera que las experiencias pueden mantener, de modo inteligible, relaciones racionales con nuestro ejercicio de la libertad implícita en la idea de espontaneidad. (McDowell 1994a: 65)

Una vez desveladas las preocupaciones que los alientan, a los afanes de la filosofía constructiva contrapone McDowell una actitud expresamente inspirada en el quietismo de Wittgenstein:

> Pues lo que yo creo que es la respuesta deberíamos ser capaces de dar, si alguien nos preguntase acerca de qué es lo que constituye la estructura del espacio de las razones, sería algo así como un encogimiento de hombros. (McDowell 1994a: 273)

Forzados, en todo caso, a esbozar una respuesta en los mismos términos en que se formula la pregunta –términos que ya habríamos impugnado–, no quedará más remedio que afirmar una cosa y la contraria, y aclarar en qué sentido se dice cada una:

> Según la concepción que he aconsejado, las capacidades conceptuales son en cierto sentido algo no natural: no podemos captar lo que significa poseer y emplear el entendimiento –una facultad de la espontaneidad– en términos de conceptos que ubiquen las cosas en el reino de la ley. Pero la espontaneidad se halla inextricablemente involucrada en la receptividad, y nuestra capacidad de receptividad –nuestros sentidos– son parte de nuestra naturaleza. De manera que, en otro sentido, las capacidades conceptuales sí que habrán de ser naturales. (McDowell 1994a: 147)

Mutatis mutandis[379]: la semántica de nuestra vida mental es algo no natural inextricablemente involucrado en la naturalidad de su sintaxis, luego también algo

[379] No es poco, dicho sea de paso, lo que debe mutarse, pues se trata de transitar del ataque de McDowell al dualismo de esquema conceptual y contenido empírico a un dualismo si se quiere menos sutil, pero al menos igual de arraigado en el seno de la psicología cognitiva, como es el dualismo entre semántica y sintaxis de las representaciones mentales, en cuyo análisis crítico venimos trabajando. Los parajes comunes entre ambos territorios intelectuales, sin embargo, deberían resultar a estas alturas suficientemente transparentes como para dispensar una elaboración más detenida.

natural. Sería, por fin, a la hora de diluir la aparente contradicción –o lo que de esa aparente contradicción persistiera tras precisar en qué sentido decimos que tal o cual aspecto de lo mental es y no es de índole natural– cuando vendría en nuestro auxilio la noción de segunda naturaleza:

> ¿Cómo podría la espontaneidad ser algo no natural –en cualquier sentido– y, empero, permanecer inextricablemente involucrada siempre que se activan nuestras actividades sensoriales?
>
> [...] Nuestra naturaleza es, en gran parte, una segunda naturaleza, y nuestra segunda naturaleza es como es debido no sólo a las potencialidades con que nacemos, sino también nuestra educación, a nuestra *Bildung*. Dada la noción de segunda naturaleza, podemos afirmar que el modo en que la razón configura nuestras vidas es natural, incluso aunque neguemos que la estructura del espacio de las razones pueda integrarse en el diseño del reino de la ley. (McDowell 1994a: 147-148)

En la noción de segunda naturaleza, de hondos ecos aristotélicos, se encuentra entonces, a ojos de McDowell, el bálsamo para esas tensiones contrapuestas que atenazan la filosofía moderna:

> El error aquí consiste en haber olvidado que la naturaleza incluye dentro de sí la *segunda naturaleza*. […L]os seres humanos adquieren una segunda naturaleza al ser adiestrados en las capacidades conceptuales, cuyas interrelaciones pertenecen al espacio lógico de las razones.
>
> Una vez que nos acordamos de que existe la segunda naturaleza, comprobamos que se pueden incluir entre las operaciones de la naturaleza ciertas circunstancias cuyas descripciones las colocan dentro del espacio lógico de las razones, por muy *sui generis* que ese espacio lógico sea. Ello hace posible que consideremos las impresiones como parte de la naturaleza sin que ello suponga una amenaza para el empirismo. (McDowell 1994a: 28)

La idea de segunda naturaleza nos franquearía el camino para que el comercio causal entre el mundo y nuestros receptores sensoriales pueda en efecto constituirse en un tribunal de jurisdicción epistémica, al quedar en alguna medida impregnado por los mismos procesos cognitivos que operan en el razonamiento. Así,

> […a] partir de la tesis de que recibir una impresión es una transacción que acaece en la naturaleza, no hay ahora ya por qué inferir […] que la idea de recibir una impresión deba resultar ajena al espacio lógico en el que funcionan conceptos como el de responsabilidad ante el mundo. Las capacidades conceptuales, cuyas interrelaciones pertenecen al *sui generis* espacio lógico de las razones, pueden resultar operativas no sólo en los juicios (que son los resultados de que un sujeto decida activamente pensar algo acerca de algo), sino que pueden serlo ya también en las transacciones naturales constituidas por los impactos por parte del mundo sobre las capacidades receptivas de un sujeto apropiado (es decir, un sujeto que posea los conceptos correspondientes). Las impresiones pueden *ser* casos en los que le aparezca perceptivamente a un sujeto (le sea manifiesto) que las cosas son de tal y cual modo. Al recibir impresiones, un sujeto puede estar abierto al modo manifiesto en que las cosas son. Esto ofrece una interpretación satisfactoria de la imagen, antes mencionada, de que hay posturas que son responsables ante el mundo a través del hecho de que son responsables ante la experiencia. (McDowell 1994a: 28-29)

Desde luego, que un fenómeno pueda ser natural sin estar integrado "en el diseño del reino de la ley" –a menos, sin estarlo en algún sentido mínimo– no dista mucho de la conclusión anhelada por los muchos intentos de afianzamiento de la autonomía de la psicología que se han sucedido desde Fodor (1968). Pero es de rigor anotar que el hecho de *llamar* "segunda naturaleza" a esa escisión parcial no calma, no mucho más que encogernos de hombros o que darle otro nombre cualquiera, nuestra necesidad de entenderla. La pregunta constructiva resurge entonces de las ascuas de lo que no hemos logrado exorcisar: ¿cómo es posible, a fin de cuentas, que tengamos tal cosa como una segunda naturaleza, esta segunda naturaleza?

Es natural, en todo caso, que cierta perplejidad no termine de quedar mitigada, pues tampoco las dificultades para engarzar nuestra simultánea pertenencia a lo que hemos dado en llamar, con Sellars (1956, *supra*), el reino de las causas y el espacio lógico de las razones dejan de recordar a lo que Lledó (2011: 9) ha descrito como el descubrimiento primordial que alienta el nacimiento, en el diálogo platónico, de "[.…] la posibilidad de entenderse a si mismo y, de paso, entender a los demás" –el descubrimiento de que:

> […h]ay dos mundos dentro del mundo de los hombres: el que vive el cuerpo, el mundo de la necesidad biológica, el mundo que transforma todo en substancia, en permanencia, lo más firme posible, dentro de la naturaleza, y el que vive la mente, el mundo de la libertad intelectual, el mundo que levanta, sobre la simple presencia de las cosas y los hombres, una luminosa y complicada constelación de deseos y sueños, de interpretaciones y sistemas conceptuales. Toda la realidad material que tocamos y que no es naturaleza ha surgido, sin embargo, como producto de ese mundo ideal que sólo existe en el lenguaje y en esa materia conformada como escultura o como vasija. (Lledó 2011: 101)

Constatar la vasta distancia entre ambos dominios, sin embargo, no basta –tampoco es bastante reiterar simplemente que, si miramos bien, son sólo uno–; no puede bastar porque, como un poco antes nos había hecho ver con delicadeza el propio Lledó (2011: 79-80):

> […] conocer no puede consistir sólo en circular, paralelamente, por cada uno de esos dos mundos. El hombre, ciudadano de ambos, está hecho de la misma materia que las cosas que percibe; no sólo *está en* la naturaleza, sino que es naturaleza. Sus sentidos pueden fraternizar con los objetos que sienten, porque sentidos y objetos son la *misma cosa*. Este mundo tiene sus leyes peculiares, la constelación de sus necesidades, sus exigencias. Nos sirve para *ser*, también, *en la naturaleza*, o sea, en la realidad, más allá de los límites del cuerpo. Este *ser naturaleza en la naturaleza* es lo que se llama *vivir*. La vida consiste en un continuo intercambio de estímulos y respuestas, consecuentes con la implacable ley de la necesidad y apoyados en la certera lógica de la *materia*, o sea, del suelo real que hermana

> e identifica el objeto con el sujeto. Pero esta doble perspectiva, esta modificación que convierte la realidad en consciencia, el fuerte mundo de los objetos en la mullida subjetividad, presenta la otra vertiente del problema. Existe la realidad palpable y su reflejo, el mundo con el que tropezamos y, dentro de las fronteras de *nuestra* naturaleza, un mundo inaprensible, un fluido incesante que sólo se remansa en el lenguaje: en unos signos sobre el papel, en las vibraciones de unas ondas sonoras que chocan y se diluyen en nuestros oídos, o brotan de nuestra boca.

Vivir: de eso trataba, después de todo, este largo estudio.

STIMULUS, MEANING, CONSCIOUSNESS: A STUDY ON THE FOUNDATIONS OF COGNITIVE PSYCHOLOGY AND THE CAUSAL EFFICACY OF THE MENTAL

Juan Hermoso Durán

Summary

This dissertation aims at unstitching some of the threads that form the question whether understanding mental life and behaviour requires a specifically psychological conceptual apparatus or, contrariwise, psychological vocabulary is a transient though long-standing tenant in the abode of science, eventually to be replaced by physiological or physical counterparts.

The attempt to apprehend the relation between states of consciousness and irritation of the nerves engendered, throughout the 19th century, a repertoire of metaphors of the unfathomable that still shapes the controversy about the mental and the physical today. The focus of the debate, however, seems to have shifted from the singular, experiential character of mental events towards their relation to whatever they refer to, thus inverting the prelation established by Émil du Bois-Reymond between the fifth and the sixth enigmas which embodied his *Ignorabimus!*, and assuming Franz Brentano's proposal to make intentionality the mark of the mental. Bertrand Russell's contention that it is suitable to think about beliefs and desires as propositional attitudes, together with Roderick M. Chisholm's effort to elucidate that territory by setting forth an analysis of the logical behaviour of the sentences used for the attribution of such attitudes in everyday language, form a particularly rich source of inspiration for the resource to internal representations in the explanation of behaviour, perhaps the main way in which cognitive psychology endeavoured to extricate itself from behaviourism. The vindication of folk psychology –the psychology implicit in such everyday talk– and the idea that the effective control of behaviour is not exerted by stimuli, but rather by their internal representations, appear, then, as deeply interrelated both in their logic and in their history. This primacy of representation is also reinforced, somewhat paradoxically, by the eviction of the notion of truth from the domain of psychology, enforced by Gottlob Frege as part of his endeavour to purge logic of the fallacies of psychologism. Some drawbacks of Chisholm's linguistic turn, as applied to our understanding of the mental, will then become apparent: it tends to depend on a biased interpretation of Fregean analyses of the notion of meaning, from which antipsychologicist overtones have been effaced, and, consequently, it tends to neglect the need to integrate an account of the possibility of erroneous reference –*i.e.*, of misrepresentation– into our theory of the nature of mental states. A brief and somewhat awkward historical excursion about the origins of our awareness of the

fissure between thoughts and things will be rehearsed in closing these introductory remarks.

The assumption that explaining intentionality entails naturalizing intentionality is deeply ingrained in contemporary debates about psychological explanation, whether that assumption is taken as a tribute to scientific reductionism or as a mere epistemological canon. Here and there, however, an inkling arises that naturalization might be an obdurate undertaking, rooted in the beguiling enticement of the assimilation of what we do not understand to what we think we do –namely, *things*–, or that intentionality might after all reveal itself as one of the landmarks signalling the territory where, as Ludwig Wittgenstein put it, "[...] explanations come to an end" –*i.e.*, as an ultimate fact. More often, nowadays, the suspicion seems to be that the *Entzauberung* of what Max Weber deemed "[…] the great enchanted garden" of the world shall finally reach every nook and cranny of the mind itself –the mind having begotten the disenchantment–, as some of the first behaviourists, like Max F. Meyer, surmised. Naturalization then, appears as a benign surrogate for antirealism about the mental, either of instrumentalist or eliminativist nuances, and the claim that the psychological vocabulary of everyday language is in fact a set of theoretical terms –hence, perhaps, embedded in a false theory, or in one which is only necessary due to a lack of more precise alternatives– takes prominence in the controversy. Once the matter is carefully weighed up, the charge of having confounded *explanans* and *explanandum* seems to disprove the antirealist arguments properly, in spite of the vague eliminativist rhetoric that pervades much recent reflection about the mind and the brain.

A firm grasp of the bonds that tie cognitive psychology to behaviourism is imperative if the prospects of an autonomous scheme of psychological explanation cast in cognitive terms are to be thoroughly assessed –a grasp, that is, beyond the customary reconstruction of behaviourism as a juvenile illness from which psychology attained little more than certain hygienic methodological habits. Quite a number of interwoven issues stand out for consideration: the fluctuating epistemological presumptions that shaped the behaviourist movement, the profuse anomalies that the theoretical structures of behaviourism were forced to absorb over the years –not any less than the elegance of some of the experimental results obtained in the laboratories of human and animal behaviour–, the often radical heterogeneity which can be found in the reflections of behaviourists themselves about the role of the mental in the explanation of behaviour, the vigorous survival of the subjective vocabulary in European psychological thought, even during the heyday of behaviourism, or the role of military research in the uprising of cognitive psychology. What will emerge from such considerations is, again, the attestation that the pretheoretical core of cognitive psychology is the idea that it is the representation of the environment, as opposed to the environment itself, that takes charge of the control of the behaviour of organisms; the fall of behaviourism, therefore, would be more exactly portrayed as a crisis of pretheoretical than as one of methodological principles.

Among the arguments that boosted the resource to theoretical constructions about internal representations, Noam Chomsky's insistence that the information present in the environment is not enough to account for our ability to learn –or develop– a language, and his devastating critique of Burrhus F. Skinner's vacuous and stealthy use of mentalistic concepts in his description of the process of linguistic learning, were to be seminal. Delving into the roots of Chomsky's theses –the validity of Markovian machines as models of human linguistic production, Karl S. Lahsley's work on the problem of serial order in behaviour, Peter Geach's and Roderick Chisholm's objections to Gilbert Ryle's attempted dispositional analysis of mental state attributions in ordinary language, William S. Verplanck's and Michael Scriven's criticisms of Skinner– will prove a fruitful route for tracking the footprints of Chomsky's thought in the development of cognitive theories. Together with positions irremediably detached from those in which cognitive psychology was moulded, it is nevertheless possible to find in behaviourism, even in the varieties of behaviourism that displayed the most entrenched hostility to theorizing about internal processes or structures, some intuitions in which the seeds of cognitive psychology are unmistakably recognizable. Thus, John B. Watson's dispute with Jacques Loeb over the epistemological autonomy of psychology –endorsed by Skinner himself–, Max F. Meyer's proposal that psychological concepts could be introduced into learning theory as abbreviations of concepts referring to complex nervous processes, and his concern, shared by Albert P. Weiss, about the need to reconcile behaviourism with some qualified version of psychophysical reductionism, his emphasis on the role that the notion of diverse criteria of classification of sensory-motor processes could play in this context, and, most importantly, Skinner's determination to provide functional definitions of any of the classes of stimuli or responses that could figure in the explanation of behaviour, are all precedents of central standpoints in cognitive psychology as clear-cut as Edward C. Tolman's or Edwin R. Guthrie's generous use of mentalistic concepts in the shape of theoretical constructs.

Both Watson and Skinner were notoriously erratic when trying to demarcate the ontological commitments entailed by their views of mind and behaviour. That certainly contributed to making Gilbert Ryle's logical behaviourism, much more perspicuous in that regard, the axis of a debate in which two themes steadily became prevailing: the categorical foundation of the behavioural dispositions that Ryle appealed to in his sketchy translations of sentences used in everyday language for the attribution of mental states, and the countless amount of caveats, alluding, precisely, to other mental states, that must be taken into account were such translations to stand a chance to hold true. A certain metaphysical repulsion for the idea that mental life might be nothing but a succession of behavioural dispositions devoid of any categorical foundation, backed by the intuition that a mental state can plausibly occur in the absence of any of the behaviours it is usually accompanied by –or even, *pace* William James, in the absence of the bodily affections that the common man would take as its expression–, an intuition brilliantly reinforced by Hilary

Putnam in a renowned *Gedankenexperiment*, was the foothold from which David M. Armstrong's contention that mental states were to be identified not with the behavioural dispositions, but with the physiological states that sustained such dispositions, would gain traction. Functionalism, called to provide the ontological bedrock for cognitive psychology, would ripen in this confluence of behaviourism and psychophysical reductionism, if only to reject both. It was, indeed, within the province of the sternest physicalism that Ullin T. Place rejected a behavioural analysis of certain mental states –raw feels– and identified them straightforwardly with states of the brain, a gesture that Armstrong widened to comprise the propositional attitudes for which Ryle had originally devised his open-ended, subjunctive conditionals. Were anything to draw states of the mind apart from irritations of the nerves, it seemed it could be nothing but the distinction between the occupant of a given causal role and the causal role itself, which, if David K. Lewis was right, might correspond to the distinction between a neurophysiological state and a psychological state.

Meanwhile, the ungovernable demeanour of mental state concepts, reappearing waywardly here and there, tacitly assumed in, or surreptitiously incorporated into dispositional analyses, had readily been denounced by Peter Geach and Roderick Chisholm, whose incisive arguments made it clear that Rylean analyses were plainly false without the *caeteris paribus* clauses in which such mental state concepts were hidden, but hopelessly vague with them, and thus provided the earliest, most general formulation of Chomsky's objections to Skinner. The task of exposing the concealed circularity of the behavioural conception of the mind aided the then incipient cognitive psychology, as Jerry Fodor has underscored, in taking charge of the relational nature of mental states. In Hilary Putnam's insightful reading of Alan M. Turing's work on automaton theory, cognitive psychology would find a means to harness that relational nature so as to avoid falling back to *virtus dormitiva* explanations: the simultaneous definition of the computational states of a system in terms of its relations to other computational states. What captured Putnam's imagination was the fact that Turing characterized logical machines as abstract automatons whose nature is defined by the machine table specifying their transition functions, quite independently of the particular physical structures in which the input, output or memory devices, or the very machine table, might happen to be instantiated. Conveyed to the relation between the mind and the brain, such a characterization of logical machines allowed Putnam to launch a formidable attack on the idea of psychophysical reduction: if the identity of mental states and states of the brain was taken as an identity between types of states, it would turn out to be as feeble as superfluous, since an equally unobjectionable naturalism followed from taking it as an identity between tokens of states, and the simple shift from the general to the particular made the thesis much less vulnerable to empirical refutation. Moreover, such a shift also implied that mental states and processes, being nothing but brain states and processes, might still, insofar as they formed kinds that could not be identified with the natural kinds delineated by neurophysiology, require an

explanatory vocabulary of their own: the vocabulary of psychology. Functionalism, then, seemed to afford naturalism without reductionism.

Notwithstanding rhetorical antagonism, there has been some controversy over whether functionalism entails a commitment to the existence of mental states and processes as such, rather than providing an effective articulation of the behaviourist project of expelling them from the vocabulary of science. A brisk look into the details of the procedure for the definition of theoretical terms contrived by Frank P. Ramsey and the notions of primary and secondary system of a scientific theory will show that, as long as that is how psychological terms are introduced into our theories, the pledge to include terms referring to mental states both as *definiendum* and as *definiens* is steadfast, and cognitive psychology maintains, in this regard, a rupture with behaviourism. It remains an open question, though, whether or not the mental states thus woven into the fabric of psychological explanation are in point of fact attributed a causal efficacy distinct from that of the brain states to which they turn out to be identical. But then again, if they are not, the yearning for epistemological autonomy for psychology will have to be quenched with the mere concession of some form of pragmatic relevance, either in the face of the hindrances of setting forth an equivalent explanation in the language of physiology, or of our bare, hopefully transient ignorance thereof.

Long before Turing hammered out the notion of an abstract automaton, the exactness of imitation shone as the greatest concern of the artisan manufacturing an automaton. Industrialization expunged this mild vanity and encouraged an effort to reproduce the essential, underlying principles of the biological or psychological phenomenon that the automaton allegedly replicated. The need to elucidate what were to be taken as essential, underlying principles led many an engineer –Silas Bent Russell, Thomas Ross, William Grey Walter, Herbert Edgard Coburn, Anthony G. Oettinger or J. Anthony Deutsch– to the conviction that the answer did not lie in details of physical structure, just as the examination of phototropic automata had taken Herbert S. Jennings to the conclusion that there are principles concerning the behaviour of organisms and machines which lie beyond the expressive resources of physics and chemistry. Both Turing himself and Clark Hull had equalled replication of essential principles to replication of behaviour, if not of physical structure, and a number of behaviourists, in the wake of Hull's work, attempted to produce robotic versions of learning processes, either Pavlovian or operant –although the robot approach awaited defiant repudiation by Skinner. Rather more taxing were the criteria proposed by Kenneth J.W. Craik, who relied on an abstract analysis of the task in hand to develop a notion of functional equivalence which would eventually prove more akin to cognitive theorizing than Turing's ingenuous behaviourism. Identity of internal processes, which had also been hinted at by Joseph Needham, Douglas G. Ellson, Hugh Bradner, Thomas Ross, Andrew G. Pask or William Gray Walter –Alan Newell, John C. Shaw, and Herbert A. Simon had audaciously reintroduced introspection to psychology by calling attention to the need to compare artificial intelligence programs with introspective protocols of mental processes–

would finally be consecrated by Fodor, on a par with the identity of possible –not only actual– behaviour, as the relevant standard for the purposes of cognitive theorizing.

Yet the upsurge of cognitive psychology found some spurious support in an aspect of Turing's work. When Turing envisioned the prospect that his logical machines, fittingly programmed, could accomplish whatever task any *computer* could accomplish, the term "computer" was used to designate human workforce: the meticulous clerk who painstakingly devoted hours and hours of labour to drudging tasks that required no ingenuity or wit beyond that of following instructions aided by pencil and paper. Any task of that kind –*i.e.*, any effective procedure– could be performed by a logical machine, Turing proved; the reconstruction of the notion of algorithm that Alonzo Church had achieved shortly before was actually equivalent to the thesis that there is a universal logical machine that can simulate any given logical machine. A mystified reading of the Church-Turing thesis to the extent that it presumably proved that Turing's machines could carry out any task whatsoever, stemming perhaps from some passing remarks by Warren S. McCullouch and Walter Pitts, gradually took shape among artificial intelligence researchers and theorists, as well as among psychologists and philosophers, and inconspicuously fostered the cognitive sciences with the force of a vision dating back to Leibniz and Lullus, the *characteristica universalis*.

This search for a perfect language is also profoundly ingrained in the foundational crisis that was shaking mathematics since the middle of the 19th century, from which the notion of thinking machines arose. As is well-known, the application to the *Entscheidungsproblem* that Turing uncovered in his logical machines aimed at addressing a challenge put forward by David Hilbert, as a response to that crisis: the foundation of all mathematical knowledge in a finite, complete, and consistent set of axioms, so that the truth of any mathematical statement formally expressed in the same terms could be effectively –*i.e.*, algorithmically– decided. The conception of scientific theories that thrived in the fertile soil of Hilbert's endeavour had a significant bearing in Putnam's persuasion that one and the same theory could be true –in fact, at some level of abstraction, ought to be true– of different organisms as well as of machines, and therefore played a prominent role in shaping the foundations of cognitive psychology. It also opened a pungent question: whether the affable metaphor of the computer of the mind that soon pervaded the work of psychologists was actually a metaphor at all, or, in other words, whether embracing cognitive psychology obliged one, as Zenon W. Pylyshyn would have it, to accept that the mind literally is a computer, and that a suitably programmed computer is a mind. Before that idea could ripen, however, the notion had to unfold that a computer's program, or a machine's blueprint, might be equivalent to a theory of the behaviour that the machine rightfully simulates –and before that notion appeared explicitly in the work of George A. Miller, Eugene Galanter, and Karl H. Pribram, it had slowly ripened in the reflections of many pioneers of automaton construction about their tireless daily efforts. The assimilation of mental and computational

processes, at any rate, has instigated some of the harshest criticisms that cognitive psychology has undergone: its contextualization as part of the dehumanizing ideology that lies beneath the brutality of the political history of the last century.

At the heart of Putnam's arguments for functionalism lies the thesis of the multiple realizability of mental states, which bestows on the critique of reductionism a stronghold for the claim that identity between types of states and processes is too severe a benchmark for a sensible account of the relation between mind and brain. The reasons that Putnam provides to take up functionalism, though, are much weaker than those urging the abandonment of reductionism: because they involve the manner in which we ordinarily identify mental states in ourselves or others, they invite backing into a behaviourist analysis; in an attempt to block that move, Putnam introduces a distinction between psychological concepts and the properties they design, but that is a distinction that a reductionist like Lewis could easily put to work against Putnam's objections to his thesis. It is only natural, then, that the exact relation between functionalism and reductionism should remain a controversial point: whereas analytical functionalism tends to combine a stalwart defence of the prospects of reductionism with the contention that the functional roles isolated through the scrutiny of everyday psychological concepts will converge with the occupants of those roles detected in neurophysiological research, empirical functionalism will usually be tied to sheer anti-reductionist positions and look for its functional roles in the empirical discoveries of psychology rather than in ordinary language. What the dispute between Putnam and Lewis boils down to, then, is whether types of mental states are to be identified with types of physical states occupying a given causal role or with sets of physically heterogeneous states sharing the functional property of occupying the characteristic functional role of that type of mental states.

In sum and substance, Putnam's position has in its favour the elegance with which it can manage the diverse incarnations that mental states of one and the same kind seem to be able to adopt; while Lewis must be credited with the straightforwardness his thesis boasts when it comes to dealing with the causal efficacy of mental states. The difficulties that Lewis' strategy is bound to face were diligently recounted by Jerry A. Fodor and Ned J. Block, and are toughened when we consider the complete loss of nomothetic scope that its proposed identity statements would suffer under a most plausible radicalization of the notion of multiple realizability –that in which it is applied not to different species or kinds of physical systems, but to different organisms of the same species, or even to one organism in different moments. Equally arduous, at any rate, are the troubles that an anti-reductionist functionalism will encounter when the need arises to account for the causal efficacy of mental states *qua* mental states, which is supposed to sustain the generalizations that the functional taxonomies of psychology afford but the vocabulary of neurophysiology presumably misses.

Despite all the noise about the metaphysical interpretation of the functionalist conception of mind, a distinctive trait of functionalism is its thorough ontological

neutrality: the pattern of functional relations that allegedly provides the fingerprint of a type of mental states could in principle be held by spiritual states of some sort of incorporeal entity –*spook stuff*, as a sardonic naturalist would have it– just as well as by physical states of chunks of matter. As much as it might seem to drive functionalism away from reductionism, it will be found that, as a matter of fact, this neutrality springs partly from John C. Smart's proposal of topic-neutral analysis as a means to elude Max Black's early objections to psychophysical reductionism.

An examination of the concepts of functional role, functional role occupant, and functional property, linked to the notions of type and token of a given state or process, will show, along the lines drawn by Stephen R. Schiffer, that empirical functionalism is committed to a theory about the propositional attitudes according to which the proposition serves as an index for the functional role of a given type of internal states whose tokens display the functional property that the propositional attitude identifies, and so to some sort of realism about mental representation. It is customary to distinguish, within the realm of empirical functionalism, between the early machine-table functionalism of Putnam's first approaches to the import of Turing's work for our understanding of the mental, Fodor's computational functionalism, marked by the introduction of a stronger notion of functional equivalence, and the homuncular functionalism of Daniel C. Dennett or William G. Lycan, in which the spotlight turns to hierarchical structures, flow diagrams and subroutine trees so as to sustain a proposal that the attribution of mental states to any of the homunculi appearing in our model of a psychological process is just as justified as to the system as a whole. Realism about mental representation, then, is mitigated in homuncular functionalism into some sort of instrumentalism.

The incessant discoveries of neurophysiological laboratories arouse a constant exertion to ensure a ruthless reduction of psychological theory to the theoretical vocabulary of brain science, an undertaking whose most vigorous conceptual upholder has been Jaegwon Kim. The crux of Kim's reproof of anti-reductionism rests upon a principle of causal inheritance, purportedly intrinsic to the realization relation, to the effect that the causal properties of the classes of which such a relation is predicated are identical. Taken at face value, this principle makes anti-reductionism an inexistent territory only appearing to lie amidst reductionism and eliminativism; Block's attempt to restore the plausibility of an anti-reductionist reading of functionalism, then, will be forced to distinguish between the idea that the defining properties of nomologically coextensive classes are projectable in terms of justification, which is rejected, and the idea that they are projectable in terms of objective evidence, which is admitted, but leaves Kim's argument at the mercy of a veiled commitment to some Lockean conception of natural kinds as nominal essences which, as Richard Boyd showed, turns out to be rather harmful to reductionism. Neither submitting to a version of the reductionist endeavour in which the scheme of psychophysical reduction is restricted to the domain of the species under study, nor trying to condemn to elimination any psychological concept which might pretend to reach beyond the confines of one species, will prove promissory for the reductionist,

since, *inter alia*, neither provides any shelter from radical varieties of multiple realizability.

The controversy regarding causal inheritance, anyhow, has over the years taken quite an unusual focus: the parallelism between the case of pain –*i.e.*, of its relation to physical states– and jade –*i.e.*, of its relation to jadeite and nephrite. It is Kim's contention, simply put, that the psychology of pain, as opposed to the neurophysiology of pain, is as infertile a scientific task as the mineralogy of jade, as opposed to that of jadeite or nephrite. But this analogy, Fodor protests, is spurious: jade is the concept of a disjunctive class, while pain is the concept of a multiply realized class; the properties of tokens of jade are therefore non-projectable and non-nomological, while the properties of tokens of pain are projectable and nomological, provided only that the appropriate theoretical vocabulary –namely, psychological vocabulary– is used to identify them. The case for Fodor's *distinguo*, though, is weak, and it will not bear much more fruit than the humble and pretty idle conclusion that the identity between pain and its functional specification happens to be closer to the core of our worldview than the identity between pain and its physical realizations, and closer, too, than the identity between jade and the aforementioned pair of silicates. A distinction between closed disjunctions like the concept of jade and open disjunctions like the concept of pain would seem to be more auspicious on Fodor's part, because it succeeds in linking the argument to a general feature of inductive reasoning: the fact that open disjunctions incite the search for laws that can explain them away through an appeal to higher-order properties. Moreover, as Block has pointed out, it is to be expected of a reducing theory that it not only replicates the generalizations entailed by the laws of the reduced theory but also explains these reduced laws, and this will not happen if the reduction depends upon open disjunctions. But the crucial drawback of Kim's approach, if Block is on the right track, is that it obliterates the distinction between the realization properties of a mental state –innocuous peculiarities of any given physical form it might adopt– and its design properties –enforced by physical and evolutionary restrictions. Or else, as a last resource perhaps, the anti-reductionist can always submit that the concept of jade might after all have a role to play in some modest branch of special science, allowing for generalizations that the disjunction of jadeite and nephrite would not licence, and thus dismantle Kim's intended analogy with the sterility of the concept of pain.

In a spirit akin to Kim's, William Bechtel and Jennifer Mundale have argued that anti-reductionism depends on a specious use of the notions of mental state and brain state, unconnected to actual scientific practice, as well as on a double standard regarding the preciseness that is required from neurophysiological and psychological kinds, and on an unjustified function-structure dichotomization of the explanatory levels that are taken as admissible in scientific theorizing. True but irrelevant, question-begging, and true and valuable, but congruent with the gist of a functionalist defence of the autonomy of psychology, will be sustained as the appraisal of each of the charges. Somewhat sterner is the upshot for John Bickle's

arguments, alternating between *petition* and *ignoratio elenchi*. The ineptness of these pretended rebuttals becomes easier to understand in the light of Carl Gillet's claim that they assume a flat conception of the realization relation –*i.e.*, one in which the causal powers of the instantiated property are a subset of those of the instantiating property, and they all must therefore be properties of the same individual–, whereas anti-reductionist functionalism is assembled on a dimensioned conception of realization –one in which the above requirements do not hold. A slight qualification is due on Mark B. Couch's distinction between two lines of anti-reductionist reasoning: it must be noted that were physicalist arguments to succeed in showing that some of the functional properties of different systems or organisms are not in fact identical, that would not provide an argument for reductionism –but for eliminativism–, unless it was accompanied by a convincing case to the effect that the physical properties underlying some other functional properties of different systems or organisms are in fact identical.

Difficult as it may be for reductionist schemes aiming at a reconstruction of the vocabulary of psychology to take hold of the intricacies of the realization relation, it has certainly proved not any less thorny for antireductionism to provide a sound account of the autonomy of psychological explanation which does not boil down to some sort of pragmatic relevance. The most spirited attempts follow the steps of Donald Davidson's insistence that our taxonomic practices as far as mental states are concerned rely on normative, rational criteria –such as the principle of charity– that do not have a place in the theoretical vocabulary of more basic sciences. Under such taxonomy, Davidson argues, mental states become involved in nomic regularities that cannot be isolated in physical or neurophysiological terms. But, of course, if Kim is right, these nomic regularities can only depend on the causal powers of the physical states that each mental state in our taxonomy happens to supervene on. This is then, what the question comes to: we want to know if our reasons for action are genuine causes of behaviour. It is important to notice, in this regard, that Davidson's proposal impugns not only Kim's principle of causal inheritance but also rather more deeply entrenched facets of our conception of causality such as the principle of causal exclusion. The conciliation of the relational nature and the causal nature of the mental will only be mature, it seems, if it comes in the wake of a profound reconsideration of our conception of causality –which might include the need to allow for different reconstructions of the relations between causes and effects to give rise to different generalizations about the world.

The path opened by Davidson has been explored by Zenon W. Pylyshyn: scientific disciplines are not demarcated by their objects or the contents of their statements, but by the conceptual properties of their chosen vocabulary; these properties delimit the *explanandum* of the theory and arrange phenomena into certain taxonomies; some taxonomies –and not others– allow for the detection of certain generalizations; the vocabulary of psychology throws light on generalizations that would remain hidden under the vocabulary of neurophysiology because the taxonomies it provides take into account properties of stimuli, internal states or

behaviours as subjectively interpreted; it is, in sum, the intentional load of the concepts of psychology that its autonomy rests upon. A look into the distinction between the extensional character of description and the intensional character of explanation that is central to Pylyshyn's proposal will once more lead to the conclusion that some reappraisal of our conception of causality is in order.

It should not be hard to recognize in Pylyshyn's thought the powerful mark of Chomsky's arguments against Skinner, which can also be clearly seen in the much earlier work of Miller, Galanter, and Pribram. More surprisingly, perhaps, it will be argued that Skinner's own explanatory project does not fit well with the cinematic description of behaviour that Pylyshyn associates with behaviourism, and that in some regards it actually anticipates some nuclear traits of the conception of psychological explanation that cognitive psychologists have bolstered. As a matter of fact, a good deal of those traits are also to be found in the classical distinction between *Erklären* and *Verstehen*, dating back at least to Johann Gustav Droysen and his 1858 analysis of the differences between *Naturwissenschaften* and *Geisteswissenshaften*, and, somewhat more transparently, in George Henrik von Wright's reconstruction of that distinction it terms of the ineliminability of teleological explanation. Both the notion that different taxonomies will provide different generalizations and the conclusion that examination of this unyielding fact will eventually force us to reconsider our conception of causality will come across the reader of von Wright. As it turns out, then, the cognitive movement in psychology is appropriately seen as a philosophy of science whose main concern is the autonomy of psychological explanation, much as behaviourism portrayed itself –or so did Skinner– as a philosophy of science, primarily concerned, though, with establishing its scientific adequacy.

Functionalism has been very reluctant to apply to stimuli and responses the procedures of functional definition that it advises for mental states. The roots of this reticence can be traced back to an effort to avoid the threat of circularity that loomed large for logical behaviourism –a concern shared by functionalists and psychophysical identity theorists, as a detailed reading of Armstrong and Block will show–, as well as to the disquieting perspective of a characterization of mental life that, totally devoid of steadfast links to reality, might, so to speak, run wild and end up being a description of just about anything at all –a hazard felicitously mapped by Block in terms of the dilemma between the chauvinism that materialists seemed helpless to avoid and the liberalism that behaviourists would be condemned to fall prey to. It will be argued, however, that steering clear of the pitfalls of chauvinism and liberalism is equally arduous for a functionalist conception of mental states whether or not it allows for a functionalist definition of stimuli and responses. But not only is the attempt to shun such difficulties futile, it is also onerous: trying to expel references to other stimuli, behaviours, or inner states from the characterization of stimuli and behaviour leads to the ostracized references reappearing surreptitiously one way or another, just as they did in the dispositional analyses devised by logical behaviourists, when they were banished from the definition of

mental states. Moreover, recognition of the functional nature of the notions of stimuli and response employed in psychological explanation would strengthen the repertoire of arguments against the reduction of psychology –it is an important part, for instance, of Fodor's case for the autonomy of the special sciences.

Two cardinal traits of the legacy of behaviourism converge in this wariness about admitting functional vocabulary into the characterization of stimuli and responses: Watson's bold rupture with the tradition of subjective, introspective reports about mental life accompanying objective reports about behaviour –Conwy Lloyd Morgan's and Thorndike's *double inference*–, and Skinner's determination to redeem the law of effect from the accusation of subjectivity held by Watson, thus unveiling for scientific study –by means of, precisely, functional definition– the realm of operant conditioning. In the territory where these two forces clash we will hear the voices of Michael Scriven, William S. Verplanck, William K. Estes, Sigmund Koch, Kenneth MacCorquodale, and Paul E. Meehl, which, just as Geach's and Chisholm's arguments against Ryle's dispositional analyses and Lashley's discussion of the problem of the serial order of behaviour, foreshadow Chomsky's critique of Skinner's project. The role of "A Review of B.F. Skinner's *Verbal Behavior*" as a foundational myth for cognitive psychology might then help explain this caginess of functionalism as regards the functional definition of stimuli and responses.

On the other hand, accepting that the description of stimuli and behaviours be constrained to a strictly physical vocabulary is, inasmuch as it takes the nervous activity of transducers and effectors as its referent, tantamount to pledging allegiance to an internalist conception of psychological explanation –one in which only the properties of our representations of things, and never the properties of things themselves, are to intervene. This in turn, as Dennett has shown, entails a view of the mind, or the brain for that matter, as a syntactic machine the semantics of whose states is entirely irrelevant to its functioning. Proximal or distal stimulation, supervenient or not-in-the-head facets of meaning, *Sinn* or *Bedeutung*: Block's distinction between narrow and broad semantic content somehow has become the pole around which the debate on the proper characterization of the semantics of mental states crystallizes. The entrenched intuition that our beliefs and desires are certainly not about such things as the electrical responses of our retinal cells –unless, of course, they just happen to be– is indeed an obstacle to any theory of the semantics of mental states in which these are to be characterized in terms of proximal stimulation. Distal stimulation, though, whether it is taken as the things in our environment themselves, as their phenomenological properties, or as the disjunction formed by them and their epistemological counterparts in other possible worlds, is a problematic candidate too –at least if Michael McDermott's arguments are sound–, and the opacity of the contents of our *de re* propositional attitudes is one of their well-known peculiarities, so proximal stimulation might not be such a bad idea after all. Assuming a functional description of stimuli and responses may nevertheless provide an escape from McDermott's conclusions, and one which is reminiscent of Pylyshyn's reflections on cognitive architecture as well as compatible with the

teleological reading of functionalism in which many, like Bechtel, have discerned a way to sidestep chauvinism and liberalism at once: the veto on functional description could advantageously be restricted to theories about early perceptual processing and motor control, while the rest of psychology is set free from such impairing constrictions and still pinned down by its links to such theories.

A determined, well-known defence of the role of broad semantic content in psychological explanations has been undertaken by Tyler Burge. Differences in mundane facts entailing differences in mental life without the mediation of changes in the state of the organism are commonplace in actual psychological explanation and represent no menace to materialism –Burge argues– unless we ignore the distinction between causal relations, or composition relations, on the one hand, and individuation relations, on the other. The blurring of a different distinction –that between properties of our mental states and properties of the linguistic expressions we use for their attribution– is according to Pierre Jacob what gives Burge's reasonings their intuitive but deceitful appeal, but a psychology assembled under such a confusion would be unable to account for the fact that we often hold mistaken beliefs about the world, since it would have the worldly facts themselves operating as constituents of our beliefs, and there is no such thing as a mistaken fact. The problem of error –*i.e.*, the need to provide an account of its possibility– is, quite undeniably, a strong motivation for internalism. But it is not only language –the vernacular of psychological explanation, in which our commonsense conception of the mind is condensed–, but history as well –the history of the things we desire or about which we hold certain beliefs– that seem to invest them with some power to affect our mental life even though we have undergone no physical change whatsoever. To the extent that cognitive psychology keeps its internalist vows, then, this would enforce a divergence between its conceptual apparatus and that of everyday, folk psychology that Lynn Rudder Baker considers reason enough to abandon functionalism –a rather hasty conclusion, we will see, since it rests upon a fallacious argument against a straw man.

The boundless diversity of beliefs and desires that we can harbour, together with the fact that holding some of them would require from us at least the ability to hold some others –*i.e.*, the productivity and systematicity that propositional attitudes share with language–, have been among the most impelling reasons to think about them in terms of their relations to language. The hypothesis that for us to believe or desire something might turn out to be one and the same thing as to maintain a certain relation to some formula of a *lingua mentis*, an inner code in which the computations postulated by psychological theory were to be developed, has also been fostered by the plausibility of assimilating the opacity of the mental to the intensionality of some linguistic expressions, or its subjectivity to the role of indexical elements in language, as well as by the indubitable advantages that the notion of a sentence in a language of thought offers over that of a proposition. Thinking about the mind under the prism of language, of course, makes it only more peremptory to face the question of the task that the semantic content of mental states is to be

assigned in psychological explanation. The formality condition that Fodor imposed on his conception of the language of thought made it patent that if syntax was to act as the link between semantics and causality, the causal efficacy of semantics would at best remain vicarious, and the metaphor of the lock and key –their shapes being bracketed with the syntax of the mechanism– provided a vigorous intuitive thrust to the hushed repudiation of semantics. The notions of syntax and semantics that had flourished in the field of proof theory and the notion of causality that naturally underlay the engineering of machines that could imitate the behaviour of organisms amalgamated in automaton theory: quite some time before Fodor coined the formality condition, the same idea could easily be found in the works of Newell, Simon and Shaw, or Minsky, and its counterpart in terms of the relation between teleological and mechanical explanation is widespread in the reflections of Loeb and Jennings, Ralph B. Perry, Nicolas Rashevsky, Alfred J. Lotka and Kenneth J.W. Craik –its echo, in fact, clearly reaches into the work of Tolman and Hull. These notions of syntax and semantics were inherited by functionalism and cognitive psychology, thereby, of course, reinforcing their commitment to an internalist view of the role of semantic content in mental life. Later, increasing disappointment with the naive logicism shared by some of the early theories of cognition and perhaps most of the early attempts at the computer simulation of mental processes would vivify the vindication of a more demanding incumbency for semantics in psychological theory.

The bonds between functionalism, methodological solipsism and the formality condition have been examined by Robert van Gulick. Insofar as functionalism is tied to an internalist notion of the semantics of mental states by means of *a priori* arguments about the local character of causality and the causal inefficacy of semantics, which are proved illegitimate by the distinction between causality and individuation set forth by Burge and Davidson, the primacy of syntax becomes an ill-founded canon. The question, then, is whether or not the explanatory vocabulary of psychology entails a commitment to causally efficacious semantic properties of mental states in a sense which turns out to be irreconcilable with any deflationist ontology –epiphenomenal, instrumentalist, eliminative–, and whether or not such a commitment imposes the rejection of internalism. Treading in the footsteps of David Braddon-Mitchell and Frank Jackson –not far, though, from some of the paths that Fred Dretske and Jerry Fodor have explored–, it becomes fairly apparent, once more, that no credible, full-blown reconstruction of the causal efficacy of the semantics of mental life will be forthcoming without profound amendments to the very notions of cause and effect. Several attempts at accounting for this sovereignty of reasons over causes will be shown come down to the paucity of our understanding of causality: Braddon-Mitchell and Jackson's appeal to the distinction between the behavioural or mental effects that are akin to the functional definition of a given mental state and those that come by as mere offshoots of its physiological realization –a distinction which leads back to Chisholm's ingenious resource to deviant causal chains in the defence of the indispensability of teleological elements in the explanation of human action–, Robert Cummins' hopeful look back at the principle of multiple realizability

–which, applied to the realization of semantic properties in syntactic properties might break the yoke of syntax just as it allegedly abolished the servitude of psychological to physiological properties–, Frank Jackson and Philip Pettit's plea for the causal relevance of dispositions, *ergo* of functional properties, as distinct from that of their categorical foundations –in which the idea of multiple realizability is again decisive–, and the refinement of the notion of a *caeteris paribus* law carefully rigged by Fodor as a response to Stephen Schiffer's sharp criticisms of its role in Fodor's epistemology of the special sciences –which involves an unwavering vindication of the semantic nature of psychological laws.

Long before the notions of syntax and semantics were wrought in the forge of automaton theory, of course, they had permeated the study of natural language. It is easy enough to find semantic differences that do not condense into syntax, just as it is to find syntactic ambiguities with a semantic import, at least while our search takes place within natural language. But a straightforward argument for the efficacy of the semantics of mental states could not possibly be framed upon such findings: we should direct our search not to everyday speech, but to the rather more elusive realm of the *lingua mentis*. That, of course, is a much more arduous task, but the difficulties are partly specious: a scrutiny of the differences between the linguistic notions of syntax and semantics that impregnate our reflections about psychological explanation and those that psychology inherited from automaton theory promptly shows that the inert stillness of semantics is in fact stipulative.

All these questions might very well turn out to be the wrong questions. If only we could release them from certain ill-advised but deep-rooted presumptions –such as, if John McDowell is right, the bonds we take for granted between perception and passivity and between judgment and spontaneity–, it would become obvious that the answers must be given in not quite the same terms as those in which the questions are posed: shrugging our shoulders or, at most, crafting an explanation of the fact that a few all too natural traits of our mental life –intentionality, teleology, spontaneity, consciousness– awoke such injudicious philosophical worries. Our living at once in the world of causes and in the world of reasons, however, our ceaseless, swift passage between them, seems to demand something more, by way of explaining, than the attestation that both spaces are natural, or that reasons form a second nature in us which is both natural and non natural. For perplexity subsists after such reasonings, and the old questions, hopefully distilled of some impurities, rearise.

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z

Bibliografía

ABNEY, S. 1996. "Statistical Methods and Linguistics," en J. Klavans y P. Resnik, eds. 1996.

ACCADEMIA NAZIONALE DEI LINCEI, eds. 1956. *I Modelli nella Tecnica: atti del convegno di Venezia, 1-4 ottobre 1955*. Roma: Accademia Nazionale dei Lincei.

ACHINSTEIN, P. 1990. "The Only Game in Town." *Philosophical Studies* 58 (3): 179-201.

ADAMS, F. y AIZAWA, K. 1994. "Fodorian Semantics," en S.P. Stich y T.A. Warfield, eds. 1994a.

ADÁRRAGA, P. 1991. "El marco general de la psicología cognitiva," en C. Castilla del Pino y J.M. Ruiz Vargas, eds. 1991.

ADORNO, T.W. 1969. "Sobre la lógica de las ciencias sociales," en T.W. Adorno, K.R. Popper, R. Dahrendorf, J. Habermas, H. Albert y H. Pilot, 1969.

ADORNO, T.W., POPPER, K.R., DAHRENDORF, R., HABERMAS, J., ALBERT, H. y PILOT, H. 1969. *La disputa del positivismo en la sociología alemana*. Edición y traducción de J. Muñoz. Barcelona: Grijalbo, 1972.

AGUSTÍN DE HIPONA. *El Maestro o Sobre el lenguaje y otros textos*. Edición y traducción de A. Domínguez. Madrid: Trotta, 2003.

——. *De Trinitate / Tratado sobre la Santísima Trinidad*. Edición y traducción de L. Arias. Madrid: La Editorial Católica, 1948.

d'ALEMBERT, J. le R. y DIDEROT, D., eds. 1751. *Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, par une société des gens de lettres.* París: Briasson, David, Le Breton, Durand. Disponible *online* en R. Morrissey, ed. University of Chicago ARTFL Encyclopédie Project (Spring 2011 Edition),: URL = <http://encyclopedie.uchicago.edu/, 16 de mayo de 2011.

ÁLVAREZ, A. 2002. "Propiedades nucleares de los fenómenos mentales según Searle: intencionalidad, subjetividad, semanticidad." *Revista de Filosofía* 27 (2): 389-417.

ALMOG, J., PERRY, J. y WETTSTEIN, H.K., eds. 1989. *Themes from Kaplan*. Oxford: Oxford University Press / Clarendon Press.

ANDERSON, A.R. 1982. "Acquisition of Cognitive Skill." *Psychological Review* 89: 369-406.

——. 1983. *The Architecture of Cognition*. Cambridge, MA.: Harvard University Press.

ANGELL, J.R. 1913. "Behavior as a Category of Psychology." *Psychological Review* 20: 255-270.

ANSCOMBE, G.E.M. 1957. *Intention*. Oxford: Blackwell.

——. 1975. "Causality and Determination," en E. Sosa, ed. 1975.

ARISTÓTELES. *Acerca del alma*. Traducción de T. Calvo. Madrid: Gredos, 1978.

——. *Ética Nicomáquea*. *Ética Eudemia*. Traducción de J. Pallí. Madrid: Gredos, 1985.

——. *Tratados de Lógica*. Traducción de M. Candel. Madrid: Gredos, 1988.

——. *Poética*. Traducción de V. García. Madrid: Gredos, 1992.

——. *Metafísica*. Traducción de T. Calvo. Madrid: Gredos, 1994.

——. *Política*. Traducción de M. García. Madrid: Gredos, 2000.

ARNAULD, A. y LANCELOT, C. 1660. *Grammaire générale et raisonnée.* [Gramática de Port-Royal]. París: Pierre le Petit.

ARNAULD, A. y NICOLE, P. 1662. *Arte de pensar ò Lógica admirable* [Lógica de Port-Royal]. Traducción de M.J.Fernández. Madrid: Ángel Corradi, 1759.

ARMSTRONG, D.M. 1966. "The Nature of Mind." *Arts (Proceedings of the Sydney University Arts Association)* 3: 37–48. Reimpreso en C.V. Borst, ed. 1970.

——. 1968. *A Materialist Theory of Mind*. Londres: Routledge & Kegan Paul. Edición revisada: Londres: Routledge, 1993.

——. 1973. *Belief, Truth, and Knowledge*. Cambridge: Cambridge University Press.

ARMSTRONG, D.M. y MALCOLM, N., eds. 1984. *Consciousness and Causality: A Debate on the Nature of Mind*. Oxford: Blackwell.

AULIO GELIO. *Noches áticas*. Edición y traducción de S. López Moreda. Madrid: Akal, 2009.

AYER, A.J. 1959. *El positivismo lógico*. Traducción de L. Aldama, U. Frisch, C.N. Molina, F.M. Torner y R. Ruiz. México, D.F.: Fondo de Cultura Económica, 1965.

BAARS, B.J. 1983. "Conscious contents provide the nervous system with coherent, global information," en R.J. Davidson, G.E. Schwartz y D. Shapiro, eds. (1983).
——. 1986. *The Cognitive Revolution in Psychology*. Nueva York, NY.: The Guilford Press.
——. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.
——. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford: Oxford University Press / Clarendon Press.
BABBAGE, C. 1822. *A Letter to Sir Humphry Davy, Bart., P.R.S., on the Application of Machinery to the Purpose of Calculating and Printing Mathematical Tables*. Londres: John Booth. Reimpreso en A. Hyman, ed. 1989.
——. 1832. *On the Economy of Machines and Manufacturers*. Londres: Charles Knight.
BACH, K. 1982. "De Re Belief and Methodological Solipsism," en A. Woodfield, ed. 1982.
——. 1983. "Russell was right (almost)." *Synthese* 54: 198-207.
——. 1986. "Thought and Object: *De Re* Representations and Relations," en M. Brand y R.M. Harnish, eds. 1986.
BACON, F. 1620. *Novum Organum*. Edición de R. Frondizi, traducción de C.F. Almori. Buenos Aires: Losada, 2003.
BAERSTEIN, H.D. y HULL, C.L. 1931. "A Mechanical Model of the Conditioned Reflex." *Journal of General Psychology* 5: 99-106.
BAKER, L.R. 1985. "A Farewell to Functionalism." *Philosophical Studies* 48: 1-13. Reimpreso en S.Silvers, ed. 1989.
——. 1989. "Instrumental Intentionality." *Philosophy of Science* 56(2): 303-316. Reimpreso en S.P. Stich y T.A. Warfield, eds. 1994a.
BARKER-PLUMMER, D. 2004. "Turing Machines," en E.N. Zalta, ed. 2008.
BARNES, J. 1979. "Parmenides and the Eleatic One." *Archiv für Geschichte der Philosophie* 61: 1-21.
BARNES, J. 1982. *Los presocráticos.* Traducción de E. Martín López. Madrid: Cátedra, 1992.
BARSALOU, L. 1987. "The Instability of Graded Structure: Implications for the Nature of Concepts," en U. Neisser, ed. 1987.
BARTLETT, F.C. 1932. *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.
BARWISE, J. 1984. *The Situation in Logic I.* Technical Report CSLI-84-2. Stanford University.
BARWISE, J. y PERRY, J. 1983. *Situations and Attitudes*. Cambridge, MA.: MIT Press.
BECHTEL, W. 1985. "Realism, Instrumentalism, and the Intentional Stance." *Cognitive Science* 9: 473-497.
——. 1988. *Filosofía de la mente. Una panorámica para la ciencia cognitiva*. Traducción de L.M. Valdés Villanueva. Madrid: Tecnos, 1991.
——. 1994. "Levels of Description and Explanation in Cognitive Science." *Minds and Machines* 4: 1-25.
BECHTEL, W. y McCAULEY, R. 1999. "Heuristic Identity Theory (or Back to the Future): the Mind-Body Problem Against the Background of Research Strategies in Cognitive Neuroscience," en M. Hahn y S.C. Stoness, eds. 1999.
BECHTEL, W. y MUNDALE, J. 1999. "Multiple Realizability Revisited: Linking Cognitive and Neural States." *Philosophy of Science* 66: 175-207.
BELINCHÓN, M., RIVIÈRE, Á. e IGOA, J.M. 1992. *Psicología del Lenguaje. Investigación y teoría*. Madrid: Trotta.
BENNETT, J. 1971. *Locke, Berkeley, Hume: Central Themes.* Oxford: Oxford University Press.
BENNETT, M.R., DENNETT, D.C., HACKER, P.M.S. y SEARLE, J.R. 2007. *La naturaleza de la conciencia. Cerebro, mente y lenguaje*. Traducción de R. Filella. Barcelona: Paidós, 2008.
BENT RUSSELL, S. 1913. "A Practical Device to Simulate the Workings of Nervous Discharges." *Journal of Animal Behavior* 3: 15-35.
BERDOY, M., WEBSTER, J.P., y McDONALD, D.W. 2000. "Fatal Attraction in Rats Infected with Toxoplasma Gondii." *Proceedings: Biological Sciences* 267 (1452): 1591-1594.
BERGMANN, G. 1956. "The Contribution of J.B. Watson." *Psychological Review* 63: 265-276.
BERMÚDEZ, J.L. 1995. "Syntax, Semantics and Levels of Explanation." *Philosophical Quarterly* 45 (180): 361-367.

BEVER, T.G., FODOR, J.A. y GARRETT, M. 1968. "A Formal Limitation of Associationism," en T. R. Dixon y D. L. Horton, eds. 1968.
BIALYSTOK, E. 1997. "Anatomy of a Revolution," en D. Martel Johnson y C.E. Erneling, eds. 1997.
BICKLE, J. 1998. *Psychoneural Reduction: The New Wave*. Cambridge, MA.: MIT Press.
——. 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.
——. 2006. "Multiple Realizability," en E.N. Zalta, ed. 2008.
BIGELOW, J. y PARGETTER, R. 1990. "The Metaphysics of Causation," *Erkenntnis* 33: 89-119.
BLACK, M. 1962. *Models and Metaphors*. Ithaca, NY.: Cornell University Press.
——. 1965. *Philosophy in America*. Londres: Routledge & Kegan Paul.
BLACKMORE, S. 2008. "La conciencia no existe. Sólo hay un cuerpo moviéndose y haciendo cosas." Entrevista de V. Carbona en *El Mundo*, 2 de marzo de 2008.
BLANCO SALGUEIRO, A. 2000. "Intrinsicidad, relacionalidad y la reconstrucción del problema del individualismo." *Revista de Filosofía* XIII (23): 105-127.
——. 2001. "El funcionalismo y las ciencias cognitivas. Comentario a P.F. Martínez-Freire: 'Base empírica y teoría funcionalista en las ciencias cognitivas'." *Ágora. Papeles de Filosofía* 20(1): 105-112.
BLAKEMORE, C. y GREENFIELD, S., eds. 1987. *Mindwaves*. Oxford: Blackwell.
BLAUG, M., ed. 1991. *Pioneers on Economics, vol. 19: William Whewell (1794-1866), Dyonisius Lardner (1793-1859), Charles Babbage (1792-1871)*. Cambridge: Edward Elgar.
BLOCK, N.J. 1978. "Troubles with Functionalism," en C.W. Savage, ed. 1978. Versión revisada en N.J. Block, ed. 1980. Traducción castellana de E.Rabossi en E. Rabossi, ed. 1996.
——, ed. 1980. *Readings in the Philosophy of Psychology, vol. I*. Cambridge, MA.: MIT Press.
——. 1980a. "What is Functionalism?," en N.J. Block, ed. 1980. Reimpreso en N.J. Block 2007a.
——. 1980b. "What Intuitions about Homunculi Do Not Show." *Behavioral and Brain Sciences* 3: 425-426. Reimpreso en N.J. Block 2007a.
——. 1980c. "Are Absent Qualia Impossible?" *Philosophical Review* 89: 257-274. Reimpreso en N.J. Block 2007a.
——. 1986. "Advertisement for a Semantics for Psychology," en P.A. French, T.E. Uehling y H.K. Wettstein, eds. 1986. Reimpreso en S.P. Stich y T.A. Warfield, eds. 1994a.
——. 1990a. "Can the Mind Change the World?," en G. Boolos, ed. 1990.
——. 1990b. "Inverted Earth." *Philosophical Perspectives* 4: 53-79.
——. 1990c. "The Computer Model of the Mind," en E.E. Smith y D.N. Osherson, eds. 1990.
——. 1995a. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18: 227-247. Versión revisada en N.J. Block, O.J. Flanagan y G. Güzeldere, eds. 1997. Reimpreso en N.J. Block 2007a.
——. 1995b. "The Mind as the Software of the Brain," en E.E. Smith y D.N. Osherson, eds. 1995.
——. 1996. "Functionalism," en D.M. Borchert, ed. 1996. Versión revisada en N.J. Block 2007a.
——. 1997. "Anti-Reductionism Slaps Back." *Noûs* 31 (*Mind, Causation, World. Philosophical Perspectives* 11): 107-133. Versión revisada en URL = <http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/AntiReductionism.pdf, 6 de febrero de 2008.
——. 2003. "Do Causal Powers Drain Away?" *Philosophy and Phenomenological Research* 67: 133-150.
——. 2006. "Max Black's Objection to Mind-Body Identity", en D.W. Zimmerman, ed. 2006. Reimpreso en N.J. Block 2007a.
——. 2007a. *Consciousness, Function, and Representation (Collected Papers)*. Cambridge, MA.: MIT Press.
——. 2007b. "Remarks on Chauvinism and the Mind-Body Problem," en N.J. Block 2007a.
BLOCK, N.J., FLANAGAN, O.J. y GÜZELDERE, G., eds. 1997. *The Nature of Consciousness. Philosophical Debates*. Cambridge, MA.: MIT Press.

BLOCK, N.J. y FODOR, J.A. 1972a. "What Psychological States Are Not." *Philosophical Review* 81: 152-181. Reimpreso en N.J. Block, ed. 1980 y N.J. Block 2007a.
——. 1972b. "Cognitivism and the Analog/Digital Distinction." Manuscrito no publicado.
BLUMENTHAL, A.L. 1970. *Language and Psychology: Historical Aspects of Psycholinguistics*. Nueva York, NY.: Wiley.
BOBROW, D. y COLLINS, A., eds. 1975. *Representation and Understanding. Studies in Cognitive Science*. Nueva York, NY.: Academic Press.
BODE, B.H. 1918. "Consciousness as Behavior." *Journal of Philosophy* 15: 449-453.
BODEN, M. 1988. *Computer Models of the Mind*. Cambridge: Cambridge University Press.
——, ed. 1990. *Filosofía de la Inteligencia Artificial*. Traducción de G. Feher de la Torre. México, D.F. : Fondo de Cultura Económica, 1994.
BOGDAN, R., ed. 1986. *Belief*. Oxford: Oxford University Press / Clarendon Press.
du BOIS-REYMOND, E.H. 1872. *Über die Grenze des Naturerkennens; die sieben Welträthsel: zwei Vorträge*. Saarbrücken: Dr. Müller, 2006.
——. 1898. *Über die Grenze des Naturerkennens; die sieben Welträthsel. Zwei Vorträge von Émil du Bois-Reymond. Des ersten Vorstrages achte, der zwei Vorträge vierte Auflage*. Leipzig: Verlag von Veit.
——. 1912. "La Mettrie," en *Reden von Émil du Bois-Reymond, I. Band*. Edición de Estelle du Bois-Reymond. Leipzig: Verlag.
BOLYAI, J. 1832. *Appendix, scientiam spatii absolute veram exhibens*. Maros-Vásárhelyini: Typis Collegis Reformatorum per Josephum et Simeonem Kali de Felsö-Vist.
BOOLOS, G., ed. 1990. *Meaning and Method. Essays in Honor of Hilary Putnam*. Cambridge: Cambridge University Press.
BORCHERT, D.M., ed. 1996. *The Encyclopedia of Philosophy Supplement*. Nueva York: Macmillan.
BORING, E.G. 1946. "Mind and Mechanism." *American Journal of Psychology* 59: 173-192.
——. 1950. *Historia de la psicología experimental*. Traducción de R. Ardila. México, D.F.: Trillas, 1978 / 1990.
——. 1959. "L'encyclopédie au Koch. Review of *Psychology: A study of a science, vol. 1.*" *Contemporary Psychology* 4: 345-346.
BORST, C.V., ed. 1970. *The Mind / Brain Identity Theory*. Londres: Macmillan.
BOWDEN, B.V., ed. 1953. *Faster than Thought (A Symposium on Digital Computing Machines)*. Londres: Sir Isaac Pitman & Sons.
BOYD, R. 1979. "Metaphor and Theory Change," en A. Ortony, ed. 1979.
——. 1980. "Materialism without Reductionism: What Physicalism Does not Entail," en N.J. Block, ed. 1980.
BRADDON-MITCHELL, D. y JACKSON, F. 1996. *Philosophy of Mind and Cognition.* Oxford: Blackwell.
BRADLEY, M.C. 1964. "Smart, J.C.C.: *Philosophy and Scientific Realism*." *Australasian Journal of Philosophy* 42: 262-283.
BRADNER, H. 1937. "A New Mechanical 'Learner'." *Journal of General Psychology* 17: 414-419.
BRAND, M. y HARNISH, R.M., eds. 1986. *The Representation of Knowledge and Belief*. Tucson, AZ.: University of Arizona Press.
BRANDOM, R. 1998. *Making it Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA.: Harvard University Press.
BRANDON, R.N. 1981. "Biological teleology: questions and explanations." *Studies in the History and Philosopy of Science* 12: 91-105.
BRANQUINHO, J., ed. 2001. *The Foundations of Cognitive Science*. Oxford: Oxford University Press / Clarendon Press.
BRECHT, B. 1927. *Bertolt Brechts Hauspostille. Mit Anleitungen, Gesangsnoten und einem Anhange*. Berlin: Propyläen-Verlag. La traducción citada del poema "Erinnerung an die Marie A." es de J. López Pacheco, en B. Brecht. *Poemas y canciones*. Traducción de J. López Pacheco y V. Romano. Madrid: Alianza, 1998.
BRELAND, K. y BRELAND, M.A. 1951. "A Field of Applied Animal Psychology." *American Psychologist* 6: 202-204.

——. 1961. "The Misbehavior of Organisms." *American Psychologist* 16: 681-684.
BRENTANO, F. 1862. *Sobre los múltiples significados del ente según Aristóteles*. Traducción de M. Abella. Madrid: Encuentro, 2007.
——. 1874. *Psicología*. Traducción de J. Gaos. Madrid: Revista de Occidente, 1926.
BREWER, W.F. 1974. "There is No Convincing Evidence for Operant or Classical Conditioning in Adult Humans," en W.B. Weimer y D.S. Palermo, eds. 1974.
BRIDGMAN, P.W. 1927. *The Logic of Modern Physics*. Nueva York, NY.: Macmillan.
BRINTON, C. 1938. *The Anatomy of Revolution.* Segunda edición revisada: 1965. Nueva York, NY.: Vintage.
BROADBENT, D. 1957. "A Mechanical Model for Human Attention and Immediate Memory." *Behavioral and Brain Sciences* 7: 55-94.
——. 1958. *Perception and Communication*. Nueva York, NY.: Pergamon Press.
——. 1980. "The Minimization of Models," en A.J. Chapman y D.M. Jones, eds. 1980.
BROCK, A., LOUW, J. y van HOORT, W., eds. 2004. *Rediscovering the History of Psychology: Essays Inspired by the Work of Kurt Danziger*. Nueva York, NY.: Kluwer.
BROCKMAN, J. ed. 2008. *Edge World Question Center 2008: What Have You Changed Your Mind About? Why?* URL = < http://edge.org/questioncenter.html, 15 de enero de 2008.
BRONCANO, F., ed. 1995. *La mente humana*. Madrid: Trotta.
BRONCKART, J.P. 2002. "La explicación en psicología ante el desafío del significado." *Estudios de Psicología* 23 (3): 387-416.
BROWN, J.S. *et al*., eds. 1953. *Current Theory and Research in Motivation: a Symposium*. Lincoln, NE.: University of Nebraska Press.
BROWN, S.C., ed. 1974. *Philosophy of Psychology*. Londres: Macmillan.
BRUCE, D. 1994. "Lashley and the Problem of Serial Order." *American Psychologist* 49(2): 93-103.
BRUNER, J.S. 1990. *Actos de significado: más allá de la revolución cognitiva*. Traducción de J.C. Gómez Crespo y J.L. Linaza. Madrid: Alianza, 1990.
——. 1997. "Will Cognitive Revolutions Ever Stop?," en D. Martel Johnson y C.E. Erneling, eds. 1997.
BRUNER, J.S., GOODNOW, J.J. y AUSTIN, G.A. 1956. *A Study of Thinking*. Nueva York, NY.: Wiley.
BUDD, M. 1989. *Wittgenstein's Philosophy of Psychology*. Londres: Routledge.
BUENO, G. 1990. "Ignoramus, Ignorabimus!" *El Basilisco* (segunda época) 4: 69-88.
BUENO, S. y PEIRANO, M., eds. 2009. *El rival de Prometeo. Vidas de Autómatas Ilustres*. Madrid: Impedimenta.
BURGE, T. 1979. "Individualism and the Mental," en P.A. French, T.E. Uehling y H.K. Wettstein, eds. 1979.
——. 1982. "Other Bodies," en A. Woodfield, ed. 1982.
——. 1986. "Individualism and Psychology." *Philosophical Review* 95: 3-46. Reimpreso en S. Silvers, ed. 1989.
BURNHAM, J.C. 1968. "On the Origins of Behaviorism," *Journal of the History of Behavioral Science* 4: 143-151.
BUTLER, J. 1726. *Fifteen Sermons Preached at the Rolls Chapel Upon the following Subjects. Upon Humane Nature. Upon the Government of the Tongue. Upon Compassion. Upon the Character of Balaam. Upon Resentment. Upon Forgiveness of Injuries. Upon Self-Deceit. Upon the Love of our Neighbour. Upon the Love of God. Upon the Ignorance of Man.* Londres: James and John Knapton. Reimpreso en *The Works of Joseph Butler: Containing The Analogy of Religion, and Sixteen Celebrated Sermons*. Londres: Adamant, 2000; facsímil de la edición de Londres: William Tegg, 1863.
BUTLER, R.J., ed. 1963. *Analytical Philosophy, vol. II*. Oxford: Basil Blackwell.
CABANIS, J.P.G. 1802. *Rapports du physique et du moral de l'homme.* París: Baillière.
CALINGER, R. 1996. *Vita Mathematica: Historical Research and Integration with Teaching*. Cambridge: Cambridge University Press.
CALVERT, G.A., SPENCE, C., y STEIN, B.E., eds. 2004. *The Handbook of Multisensory Processes*. Cambridge, MA.: MIT Press.

CAMPBELL, J. 2002. *Reference and Consciousness*. Oxford: Oxford University Press / Clarendon Press.
CAMPOS-ROLDÁN, M. 1999. "Balance y liquidación del conductismo." *Revista de Psicología* 5: 77-112.
CANTOR, G. 1891. "Über eine elementare Frage der Mannigfaltigkeitslehre." *Jahresbericht der Deustchen Mathematikher Vereinigung* 1: 75-78.
CAPITAN, W.H. y MERRILL, D.D., eds. 1967. *Art, Mind, and Religion*. Pittsburgh, PA.: University of Pittsburgh Press.
CARNAP, R. 1928. *La construcción lógica del mundo*. Traducción de L. Mues de Schrenk. México, D.F.: Universidad Nacional Autónoma, 1988.
——. 1932. "Psychologie in physikalischer Sprache." *Erkenntnis* 3: 107-142. Reimpreso como "Psychology in Physicalist Language" en A.J. Ayer, ed. 1959. Traducción de G. Schick.
——. 1938. "Logical Foundations of the Unity of Science," en *International Encylopedia of Unified Science*, I, 1: 42-62. Chicago, IL.: University of Chicago Press.
——. 1956. "The Methodological Character of Theoretical Concepts," en H. Feigl y M. Scriven, eds. 1956.
——. 1963. "Herbert Feigl on Physicalism," en P.A. Schilpp, ed. 1963.
CARPENTER, B.E., y DORAN, R.W., eds. 1986. *A.M. Turing's ACE Report of 1946 and Other Papers*. Cambridge, MA.: MIT Press.
CARRITT, E.F. 1947. *Ethical and Political Thinking*. Oxford: Oxford University Press / Clarendon Press.
CARRUTHERS, P. y SMITH, P.K., eds. 1996. *Theories of Theories of Mind*. Cambridge: Cambridge University Press.
CARRUTHERS, P. 2005. *Consciousness. Essays from a Higher-Order Perspective.* Oxford: Oxford University Press / Clarendon Press.
CASTAÑEDA, H.N., ed. 1967. *Intentionality, Mind, and Perception*. Detroit, MI: Wayne State University Press.
CASTILLA DEL PINO, C. y RUIZ VARGAS, J.M., eds. 1991. *Aspectos cognitivos de la esquizofrenia*. Madrid: Trotta.
CHACÓN, P. *et al.* 2001. *Filosofía de la Psicología*. Madrid: Biblioteca Nueva.
CHACÓN, P. y HERMOSO, J. 2009. "New Work on the Ontology of the Unconscious: Things We've Learnt Against Searle," en L. Fernández Moreno, ed. 2009.
CHACÓN, P. y RODRÍGUEZ, M., eds. 2000. *Pensando la mente. Perspectivas en Filosofía y Psicología*. Madrid: Biblioteca Nueva.
CHACÓN, P. y RODRÍGUEZ, M. 2001. "Fisicalismos," en P. Chacón *et al.* 2001.
CHALMERS, D. 1996. *La mente consciente: en busca de una teoría fundamental.* Traducción de J.A. Álvarez. Barcelona: Gedisa, 1996.
CHAPMAN, A.J. y JONES, D.M., eds. 1980. *Models of Man*. Londres: British Psychological Society.
CHASE, W., ed. 1973. *Visual Information Processing*. Nueva York, NY.: Academic Press.
CHIHARA, C. y FODOR, J.A. 1965. "Operationalism and Ordinary Language." *American Philosophical Quarterly* 2(4): 281-295.
CHISHOLM, R.M. 1956. "Sentences about Believing." *Proceedings of the Aristotelian Society* 56: 125-148.
——. 1957. *Perceiving: A Philosophical Study*. Ithaca, NY.: Cornell University Press.
——. 1958. "Sentences about Believing," en H. Feigl, M. Scriven y G. Maxwell, eds. 1958.
——. 1966. "Freedom and Action," en K. Lehrer, ed. 1966.
——. 1967. "Intentionality," en E. Edwards, ed. 1967.
——. 1980. "The Logic of Believing." *Pacific Philosophical Quarterly* 61: 31-49.
——. 1991. "Intentional Inexistence," en D.M. Rosenthal, ed. 1991. Reimpresión del capítulo XI de R.M. Chisholm 1957.
CHOMSKY, N. 1956. "Three Models for the Description of Language." *IRE Transactions on Information Theory* IT-2 (3): 113-124.
——. 1957. *Syntactic Structures*. La Haya: Mouton.

——. 1959. "A Review of B.F. Skinner's *Verbal Behavior*." *Language* 35(1): 26-58. Reimpreso en H. Geirsson y M. Losonsky, eds. 1996.
——. 1966. *Topics in the Theory of Generative Grammar*. La Haya: Mouton.
——. 1968. *Language and Mind*. Nueva York, NY.: Harcourt, Brace & World.
——. 1976. *Reflections on Language*. Londres: Fontana / Collins.
——. 1980. *Rules and Representations*. Nueva York, NY.: Columbia University Press.
——. 1986a. "Changing Perspectives on Knowledge and Use of Language," en M. Brand y R.M. Harnish, eds. 1986.
——. 1986b. "Interview with Noam Chomsky," en B.J. Baars 1986.
——. 1993. "A Minimalist Program for Linguistic Theory". *MIT Occasional Papers in Linguistics* 1: 1-67. Reimpreso en N. Chomsky, 1995.
——. 1995. *El programa minimalista*. Traducción de J. Romero. Madrid: Alianza, 1999.
CHOMSKY, N. y MILLER, G.A. 1958. "Finite-State Languages." *Information and Control* 1: 91-112.
CHURCH, A. 1932. "A set of Postulates for the Foundation of Logic." *Annals of Mathematics* (II) 33: 346-366.
——. 1936a. "An Unsolvable Problem of Elementary Number Theory." *American Journal of Mathematics* 58: 345-363.
——. 1936b. "A Note on the *Entscheidungsproblem*." *Journal of Symbolic Logic* 1: 40-41.
CHURCHLAND, P.M. 1981. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78(2): 67-90. Reimpreso en A.I. Goldman, ed. 1993.
——. 1984. *Matter and Consciousness. A Contemporary Introduction to the Philosophy of Mind*. Cambridge, MA.: MIT Press. Segunda edición revisada en 1988.
——. 1985. "Reduction, Qualia, and the Direct Introspection of Brain States." *Journal of Philosophy* 82: 8-28.
——. 1986. "Some Reductive Strategies in Cognitive Neurobiology." *Mind* 95: 279-309. Reimpreso en S. Silvers, ed. 1989.
CHURCHLAND, P.M. y CHURCHLAND, P.S. 1981. "Functionalism, Qualia, and Intentionality." *Philosophical Topics* 12: 121-145.
——. 1983. "Stalking the Wild Epistemological Engine." *Nous* 17: 5-18.
——. 1990. "Could a Machine Think?" *Scientific American* 262: 26-31.
CHURCHLAND, P.S. 1978. "Fodor on Language Learning." *Synthese* 38: 149-159.
——. 1986. *Neurophilosophy*. Cambridge, MA.: MIT Press.
CLAPIN, H., ed. 2002a. *Philosophy of Mental Representation*. Oxford: Oxford University Press / Clarendon Press.
——. 2002b. "Tacit Representation in Functional Architecture." En H. Clapin, ed. 2002a.
CLARK, A. 2002. "Minds, Brains, and Tools." En H. Clapin, ed. 2002a.
CLARK, H.H y CLARK, E.V. 1977. *Psychology and Language*. Nueva York, NY.: Harcourt Brace Jovanovich.
COBURN, H.E. 1951. "The Brain Analogy." *Psychological Review* 58: 155-178.
——. 1952. "The Brain Analogy: a Discussion." *Psychological Review* 59: 453-460.
——. 1953a. "The Brain Analogy: Association Tracts." *Psychological Review* 60: 197-206.
——. 1953b. "The Brain Analogy: Transfer of Differentiation." *Psychological Review* 60: 413-422.
COFER, C.N. y MUSGRAVE, B.S., eds. 1963. *Verbal Behavior and Learning*. Nueva York, NY.: McGraw-Hill.
COLLI, G. 1988. *La Naturaleza ama esconderse*. Traducción de M. Morey. Madrid: Siruela, 2008.
COLLINGWOOD, R.G. 1946. *Idea de la historia*. Traducción de E. O'Gorman y J. Hernández. México, D.F.: Fondo de Cultura Económica, 1952.
COMTE, A. 1830-1842. *Cours de Philosophie Positive*. París: Bachelier.
de CONDILLAC, E.B. 1754. *Tratado de las sensaciones*. Edición de R. Mondolfo. Buenos Aires: Eudeba, 1963.
COPELAND, B.J. 2002. "The Church-Turing Thesis," en E.N. Zalta, ed. 2008.
CORDESCHI, R. 2002. *The Discovery of the Artificial. Behavior, Mind and Machines Before and Beyond Cybernetics*. Dordrecht: Kluwer.

CORDESCHI, R. y TAMBURRINI, G. 2006. *Intelligent Machines and Warfare: Historical Debates and Epistemologically Motivated Concerns*. Londres: College.
CORNMAN, J. 1962. "The Identity of Mind and Body." *Journal of Philosophy* 59: 486-492.
COWAN Jr, C.L., REINES, F. y HARRISON, F.B. 1956. "Detection of the Free Neutrino: A Confirmation." *Science* 124: 103.
CRAIG, W. 1953. "On Axiomatizability Within a System." *Journal of Symbolic Logic* 18 (1): 30-32.
CRAIK, K.J.W. 1943. *The Nature of Explanation*. Cambridge: Cambridge University Press.
——. 1947. "Theory of the Human Operator in Control Systems, I. The Operator as an Engineering System." *British Journal of Psychology* 38: 56-61.
——. 1948. "Theory of the Human Operator in Control Systems, II. Man as an Element in a Control System." *British Journal of Psychology* 38: 142-148.
——. 1966. *The Nature of Psychology*. Edición a cargo de S.L. Sherwood. Cambridge: Cambridge University Press.
CRAVER, C.F. 2001. "Role Functions, Mechanisms, and Hierarchy." *Philosophy of Science* 68: 31-55.
CUENCA, M.J. y HILFERTY, J. 1999. *Introducción a la lingüística cognitiva*. Barcelona: Ariel.
CUMMINS, R. 1975. "Functional Analysis." *Journal of Philosophy* 72(20): 741-764.
——. 1983. *The Nature of Psychological Explanation*. Cambridge, MA.: MIT Press.
——. 1986. "Inexplicit Information," en M. Brand y R.M. Harnish, eds. 1986.
——. 1989. *Meaning and Mental Representation*. Cambridge, MA.: MIT Press.
——. 1996. *Representations, Targets, and Attitudes*. Cambridge, MA.: MIT Press.
——. 2000. "Reply to Millikan." *Philosophy and Phenomenological Research* 60: 113-128.
——. 2002a. "Haugeland on Representation and Intentionality," en H. Clapin, ed. 2002a.
——. 2002b. "Comments on Smith on Cummins," en H. Clapin, ed. 2002a.
DALGARNO, M. y MATHEWS, E., eds.1987. *The Philosophy of Thomas Reid*. Dordrecht: Kluwer.
DANZIGER, K. 1990. *Constructing the Subject. Historical Origins of Psychological Research*. Cambridge: Cambridge University Press.
——. 1997. *Naming the Mind. How Psychology Found its Language*. Londres: Sage.
DARWIN, C. 1859. *El origen de las especies*. Traducción de A. de Zulueta. Madrid: Espasa-Calpe, 1998.
——. 1871. *El origen del hombre*. Prólogo y traducción de J. Roc. Barcelona: Crítica, 2009.
——. 1872. *La expresión de las emociones en los animales y el hombre*. Traducción de R. Fernández Rodríguez. Madrid: Alianza, 1998.
DAVIDSON, D. 1963. "Actions, Reasons, and Causes." *Journal of Philosophy* 60: 685-700. Reimpreso en D. Davidson 1980.
——. 1966. "Emeroses by Other Names." *Journal of Philosophy* 63: 778-780.
——. 1967. "Truth and Meaning." *Synthese* 17: 304-323. Reimpreso en D. Davidson 2001.
——. 1970. "Mental Events," en L. Foster y J. W. Swanson, eds. 1970. Reimpreso en D. Davidson 1980.
——. 1973a. "The Material Mind," en P. Suppes, L. Henkin, G.C. Moisil, y A. Joja, eds. 1973. Reimpreso en D. Davidson 1980.
——. 1973b. "In Defense of Convention T," en H. Leblanc, ed. 1973. Traducción castellana: "En defensa de la Convención T", en D. Davidson 1984.
——. 1973c. "Radical Interpretation." *Dialectica* 27: 314-328. Traducción castellana: "Interpretación radical", en D. Davidson 1984.
——. 1974a. "Philosophy as Psychology," en S. C. Brown, ed. 1974. Reimpreso en D. Davidson 1980.
——. 1974b. "Belief and the Basis of Meaning."*Synthese* 27: 309-323. Traducción castellana: "La creencia y el fundamento del significado", en D. Davidson 1984.
——. 1974c. "On the very idea of a conceptual scheme". *Proceedings and Addresses of the American Philosophical Association* 47: 5-20. Traducción castellana: "De la idea misma de un esquema conceptual", en D. Davidson 1984.

——. 1980. *Ensayos sobre acciones y sucesos*. Traducción de O. Hansberg, J.A. Robles y M. Valdés. Barcelona / México, D.F.: Crítica / UNAM, 1995. Segunda edición revisada, con dos ensayos nuevos: *Essays on Actions and Events*. Oxford: Oxford University Press, 2001.
——. 1984. *De la verdad y de la interpretación: fundamentales contribuciones a la teoría del lenguaje*. Traducción de G. Filippi. Barcelona: Gedisa, 1990.
——. 1987. "Knowing One's Own Mind." *Proceedings and Addresses of the American Philosophical Association* 61: 441-458.
——. 1992. *Mente, mundo y acción*. Traducción de C.J. Moya. Barcelona: Paidós, 1992.
——. 1993. "Thinking Causes," en J. Heil y A. Mele, eds. 1993.
——. 2001. *Subjetivo, intersubjetivo, objetivo*. Traducción de O. Fernández Prat. Madrid: Cátedra, 2003.
DAVIDSON, R.J. , SCHWARTZ, G.E. y SHAPIRO, D. eds. 1983. *Consciousness and Self-Regulation*. Nueva York, NY.: Plenum Press.
DAVIS, M., ed. 1965. *The Undecidable*. Nueva York, NY.: Raven.
DENNETT, D.C. 1969. *Content and Consciousness*. Londres: Routledge & Kegan Paul.
——. 1971. "Intentional Systems." *Journal of Philosophy* 68: 87-106.
——. 1973. "Mechanism and Responsibility," en T. Honderich, ed. 1973. Reimpreso en D.C. Dennett 1978.
——. 1975. "Brain Writing and Mind Reading," en K. Gunderson, ed. 1975. Reimpreso en D.C. Dennett 1978b.
——. 1976. "Conditions of Personhood," en A.O. Rorty, ed. 1976. Reimpreso en D.C. Dennett 1978.
——. 1978a. "Toward a Cognitive Theory of Consciousness," en C.W. Savage, ed. 1978. Reimpreso en D.C. Dennett 1978.
——. 1978b. *Brainstorms. Philosophical Essays on Mind and Psychology*. Cambridge, MA.: Bradford Books.
——. 1981. "True Believers: The Intentional Strategy and Why it Works," en A.F. Heath, ed. 1981. Reimpreso en S.P Stich y T.A. Warfield, eds. 1994a.
——. 1982. "Beyond Belief," en A. Woodfield, ed. 1982.
——. 1983. "Styles of Mental Representation." *Proceedings of the Aristotelian Society* 83: 213-226. Reimpreso en D.C. Dennett 1987.
——. 1984. "Cognitive Wheels: the Frame Problem of AI," en C. Hookway, ed. 1984.
——. 1986. "The Logical Geography of Computational Approaches: A View from the East Pole," en M. Brand y R.M. Harnish, eds. 1986.
——. 1987. *La actitud intencional*. Traducción de G. Ventureira. Barcelona: Gedisa, 1991.
——. 1988. "Quining Qualia," en A. Marcel y E. Bisiach, eds. 1988.
——. 1991a. *La conciencia explicada: una teoría interdisciplinar.* Traducción de S. Balari Ravera. Barcelona: Paidós, 1995.
——. 1991b. "Granny's Campaign for Safe Science," en B. Loewer y G. Rey, eds., 1991.
——. 1995. *La peligrosa idea de Darwin: evolución y significados de la vida*. Traducción de C. Pera Blanco-Morales. Barcelona: Galaxia Gutenberg.
——. 1996. *Tipos de mente: hacia una comprensión de la conciencia*. Traducción de F. Páez. Madrid: Debate, 2000.
——. 1999. "Review of John Haugeland: *Having Thought: Essays in the Metaphysics of Mind*." *Journal of Philosophy* 96 (8): 430-435.
——. 2000. "Making Tools for Thinking," en D. Sperber, ed. 2000.
——. 2001a. "Things about Things," en J. Branquinho, ed. 2001.
——. 2001b. "Are We Explaining Consciousness Yet?" *Cognition* 79: 221-237. Reimpreso en D. Dennett 2005.
——. 2002. "Brian Cantwell Smith on Evolution, Objectivity, and Intentionality," en H. Clapin, ed. 2002a.
——. 2005. *Dulces sueños. Obstáculos filosóficos para una ciencia de la conciencia*. Traducción de J. Barbera y S. Jawerbaum. Buenos Aires: Katz, 2006.

——. 2008. "Competition in the brain," en J. Brockman, ed. *Edge World Question Center 2008: What Have You Changed Your Mind About? Why?* URL = <http://edge.org/q2008/q08_index.html#dennett, 15 de enero de 2008.
DESCARTES, R. 1628/1701. *Reglas para la dirección del espíritu*. Traducción de J.M. Navarro Cordón. Madrid: Alianza, 2010.
——. 1637. *Discurso del método*. Edición y traducción de E. Bello. Madrid: Tecnos, 2003.
——. 1641/1642. *Meditaciones metafísicas con objeciones y respuestas*. Edición y traducción de V. Peña. Madrid: Alfaguara, 1977.
——. *Meditaciones Metafísicas y otros textos*. Traducción y edición de E. López y M. Graña. Madrid: Gredos: 1987.
——. 1649. *Las pasiones del alma*. Edición de J.A. Martínez Martínez, traducción de J.A. Martínez Martínez y P. Andrade Boué. Madrid: Tecnos, 1997.
——. *Correspondencia con Isabel de Bohemia y otras cartas*. Traducción de M.T. Gallego Urrutia, introducción de M. Cabot. Barcelona: Alba, 1999.
DEUTSCH, J.A. 1953. "A New Type of Behaviour Theory." *British Journal of Psychology* 44: 304-318.
——. 1954. "A Machine with Insight." *Quarterly Journal of Experimental Psychology* 6 (1): 6-11.
——. 1955. "The Insightful Learning Machine." *Discovery* 16 (12): 514-517.
DEVITT, M. 1981. *Designation*. Cambridge: Cambridge University Press.
——. 1989. "A Narrow Representational Theory of the Mind," en S. Silvers, ed. 1989.
DEWEY, J. 1896. "The Reflex Arc Concept in Psychology." *Psychological Review* 3: 357-370.
DILTHEY, W. 1883. *Introducción a las ciencias del espíritu*. Traducción de J. Marías. Madrid: Revista de Occidente, 1966.
DIÓGENES LAERCIO. *Vidas, opiniones y sentencias de los filósofos más ilustres*. Traducción de J. Ortiz y Sanz. Madrid: Luis Navarro, 1887.
DIRICHLET, P.G.L. 1837. "Beweis des Satzes, dass jede unbegrenzte arithmetische Progression, deren erstes Glied und Differenz ganze Zahlen ohne gemeinschaftlichen Factor sind, unendlich viele Primzahlen enthält." *Abhandlungen der Königlich Preussisch Akademie der Wissenschaften* 48: 45-81.
DIXON, T.R. y HORTON, D.L., eds. 1968. *Verbal Behavior and General Behavior Theory*. Englewood Cliffs, NJ.: Prentice-Hall.
DOMÍNGUEZ, A. 2003. Introducción a Agustín de Hipona, *El Maestro o Sobre el lenguaje y otros textos*.
DONNELLAN, K. 1966. "Reference and Definite Descriptions." *Philosophical Review* 75: 281-304.
DRAAISMA, D. 1993. *Las metáforas de la memoria. Una historia de la mente*. Traducción de C. Ginard. Madrid: Alianza, 1998.
DRAY, W.H. 1957. *Laws and Explanation in History*. Oxford: Oxford University Press.
DRETSKE, F. 1980. "The Intentionality of Cognitive States," en P.A. French, T.E. Uehling y H.K. Wettstein, eds. 1980.
——. 1981. *Knowledge and the Flow of Information*. Cambridge, MA.: MIT Press.
——. 1983a. "Précis of Knowledge and the Flow of Information." *Behavioral and Brain Sciences* 6: 55-63.
——. 1983b. "The Epistemology of Belief." *Synthese* 55: 3-19.
——. 1985. "The Explanatory Role of Content," en D.D. Merrill y R.H. Grimm, eds. 1988.
——. 1986a. "Aspects of Cognitive Representation," en M. Brand y R.M. Harnish, eds. 1986.
——. 1986b. "Misrepresentation," en R. Bogdan, ed. 1986. Reimpreso en S.P. Stich y T.A. Warfield, eds. 1994a.
——. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA.: MIT Press.
——. 1989. "Reasons and Causes." *Philosophical Perspectives* 3: 1-15.
——. 1995. *Naturalizing the Mind*. Cambridge, MA.: MIT Press.
DREYFUS, H. 1972. *What Computers Can't Do*. Nueva York, NY.: Harper & Row. Segunda edición revisada: 1979. Nueva York, NY.: Harper & Row. Tercera edición revisada: 1992. *What Computers Still Can't Do*. Cambridge, MA.: MIT Press.
DREYFUS, H. y DREYFUS, S. 1985. *Mind over Machine*. Nueva York, NY.: Free Press.

——. 1988. "La construcción de una mente *versus* el modelaje del cerebro: la Inteligencia Artificial regresa a un punto de ramificación," en M. Boden, ed. 1990.
DROYSEN, J.G. 1858. *Grundriss der Historik*. Leipzig: Verlag von Veit, 1868.
DUHEM, P. 1906. *La teoría física: su objeto y estructura*. Traducción de M. Pons Irazazábal. Barcelona: Herder, 2003.
DUNCAN, D.M. 1972. "James G. Miller Living Systems Theory: Issues for Management Thought and Practice." *The Academy of Management Journal* 15 (4): 513-523.
von ECKARDT, B. 1993. *What is Cognitive Science?* Cambridge, MA.: MIT Press.
ECO, U. 1993. *La búsqueda de la lengua perfecta*. Traducción de M. Pons. Barcelona: Crítica, 1994.
EDELMAN, G. 1987. *Neural Darwinism: the Theory of Neuronal Group Selection*. Nueva York, NY.: Basic Books.
EDELMAN, G. y MOUNTCASTLE, V.B. 1978. *The Mindful Brain: Cortical Organization and the Group-Selective Theory of Higher Brain Function.* Cambridge: Cambridge University Press.
EDWARDS, E., ed. 1967. *The Encyclopedia of Philosophy*. Nueva York, NY.: Macmillan.
EGGER, M.D. y MILLER, N.E. 1962. "Secondary Reinforcement in Rats as a Function of Information Value and Reliability of the Stimulus." *Journal of Expermimental Psychology* 64: 97-104.
EGUREN, L. y SORIANO, O. 2004. *Introducción a una sintaxis minimalista*. Madrid: Gredos.
ELLSON, D.G. 1935. "A Mechanical Synthesis of Trial-and-Error Learning." *Journal of General Psychology* 13: 212-218.
ENÇ, B. 1983. "In Defense of the Identity Theory." *Journal of Philosophy* 80: 279-298.
ENDICOTT, R. 1993. "Species-Specific Properties and More Narrow Reductive Strategies." *Erkenntnis* 38: 303-321.
EPICTETO. *Enquiridión*. Traducciones de J.M. García de la Mora y F. de Quevedo. Madrid: Anthropos, 2004.
ERDELYI, M., 1985. *Psicoanálisis: la psicología cognitiva de Freud*. Traducción de N. Daurella. Barcelona: Labor, 1987.
ESCOBAR, R. y LATTAL, K.A. 2011. "Observing Ben Wyckoff: From Basic Research to Programmed Instruction and Social Issues." *Behavior Analysis* 34 (2): 149-170.
ESTANY, A. 1999. *Vida, muerte y resurrección de la conciencia. Análisis filosófico de las revoluciones científicas en la psicología contemporánea*. Barcelona: Paidós.
ESTES, W.K., KOCH, S., MACCORQUODALE, K., MEEHL, P.E., MUELLER, C.N., SHOENFELD, W.N. y VERPLANCK, W.S., eds. 1954. *Modern Learning Theory: A Critical Analysis of Five Examples*. Nueva York, NY.: Appleton-Century-Crofts.
EVANS, E.D. 2000. "J.M. Stephens (1901-2000). Obituary." *American Psychological Association Newsletter for Educational Psychologists* 24 (1): 15.
EVANS St. B.T., J. 1972. "Interpretation and Matching Bias in a Reasoning Task." *Quarterly Journal of Experimental Psychology* 24: 193-199.
FALK, W.D. 1948. "'Ought' and Motivation." *Proceedings of the Aristotelian Society* 48: 111-138.
FECHNER, G.T. 1851. *Zend-Avesta; oder Über die Dinge des Himmels und des Jenseits, von Standpunkt der Naturbetrachtung.* Leipzig: L. Voss.
FEINGENBAUM, E.A. 1959. "An Information Processing Theory of Human Verbal Learning." *The Rand Corporation Mathematics Division* P-1817.
FEIGL, H. 1958. "The 'Mental' and the 'Physical'," en H. Feigl, M. Scriven y G. Maxwell, eds. 1958. Edición revisada, con un ensayo nuevo: *The 'Mental' and the 'Physical': The Essay and a Postscript.* Minneapolis, MN.: University of Minnesota Press, 1967.
——. 1960. "The Mind-Body Problem: Not a Pseudo-Problem," en S. Hook, ed. 1960.
FEIGL, H. y MAXWELL, G. 1962. *Scientific Explanation, Space, and Time (Minnesota Studies in the Philosophy of Science III)*. Minneapolis, MN.: University of Minnesota Press.
FEIGL, H. y SCRIVEN, M., eds. 1956. *The Foundations of Science and the Concepts of Psychology and Psychoanalysis (Minnesota Studies in the Philosophy of Science I)*. Minneapolis, MN.: University of Minnesota Press.

FEIGL, H., SCRIVEN, M. y MAXWELL, G., eds. 1958. *Concepts, Theories, and the Mind-Body Problem (Minnesota Studies in the Philosophy of Science II)*. Minneapolis, MN.: University of Minnesota Press.
FELDMAN, R. 2001. "Naturalized Epistemology," en E.N. Zalta, ed. 2008.
FERNÁNDEZ MORENO, L. 2006. *La referencia de los nombres propios*. Madrid: Trotta.
——, ed. 2009. *Language, Nature, and Science: New Perspectives*. Madrid: Plaza y Valdés.
FERNÁNDEZ TRESPALACIOS, J.L. 1992. "Skinner y la psicología cognitiva," en J.A. Mora, ed. 1992a.
FEYERABEND, P. 1963. "Mental Events and the Brain." *Journal of Philosophy* 60: 295-296.
——. 1978. *Science in a Free Society*. Londres: New Left Books.
——. 2009. *Filosofía natural*. Traducción de J. Chamorro Mielke. Barcelona: Debate, 2013.
FIELD, H.H. 1977. "Logic, Meaning and Conceptual Role." *Journal of Philosophy* 74: 379-409.
——. 1978. "Mental Representation." *Erkenntnis* 13 (1): 9-61. Reimpreso en S.P. Stich y T.A. Warfield, eds. 1994a.
——. 1986. "The Deflationary Conception of Truth," en G. Macdonald y C. Wright, eds. 1987.
FILÓSTRATO, F. *Vida de Apolonio de Tiana*. Edición de A. Bernabé Pajares. Madrid: Gredos, 1992.
FITTS, P.M. y POSNER, M.I. 1967. *Learning and skilled performance in human performance*. Belmont, CA.: Brock-Cole.
FLAMMARION, N.C. 1888. *L'atmosphere : Météorologie populaire*. París: Hachette.
FLOURENS, J.M. 1824. *Recherches expérimentales sur les propriétés et les fonctions du système nerveux dans les animaux vertébrés*. París: Crevot.
——. 1842. *Examen de la phrénologie: réfutation des doctrines matérialistes de Gall, Spürhzeim et Broussais*. París: Paulin.
——. 1864. *Examen du livre du M. Darwin sur l'origine des espèces*. París: Garnier.
FERSTER, C.B. y SKINNER, B.F. 1957. *Schedules of Reinforcement*. East Norwalk, CT.: Appleton-Century-Crofts.
FESTINGER, L. 1957. *Teoría de la disonancia cognitiva*. Traducción de J.E. Martín. Madrid: Centro de Estudios Políticos y Constitucionales, 1975.
FESTINGER, L. y CARLSMITH, J.M. 1959. "Cognitive consequences of forced compliance." *Journal of Abnormal and Social Psychology* 58: 203-210.
FODOR, J.A. 1965. "Explanations in Psychology," en M. Black, ed. 1965.
——. 1968a. "The Appeal to Tacit Knowledge in Psychological Explanations." *Journal of Philosophy* 65: 627-640. Reimpreso en J.A. Fodor (1981a).
——. 1968b. *La explicación psicológica. Introducción a la filosofía de la psicología*. Traducción y notas de J.E. García Albea. Madrid: Cátedra, 1991.
——. 1974. "Special Sciences: Or the Disunity of Science as a Working Hypothesis." *Synthese* 28: 97-115.
——. 1975. *El lenguaje del pensamiento*. Traducción de J. Fernández. Madrid: Alianza, 1985.
——. 1980a. "Methodological Solipsism Considered as a Research Strategy in Cognitive Science." *Behavioral and Brain Sciences* 3: 63-73.
——. 1980b. "Fixation of Belief and Concept Acquisition," en M.Piatelli-Palmarini, ed. 1980.
——. 1981a. *RePresentations*. Cambridge, MA.: MIT Press.
——. 1981b. "The Present Status of the Innateness Controversy," en J. Fodor 1981a.
——. 1981c. "The Mind-Body Problem." *Scientific American* 244: 124-132.
——. 1983. *La modularidad de la mente. Un ensayo sobre la psicología de las facultades.* Traducción de J.E. García-Albea. Madrid: Morata, 1986.
——. 1984. "Semantics, Wisconsin Style." *Synthese* 59: 231-250. Reimpreso en S. Silvers, ed. 1989 [y en J.A. Fodor 1990].
——. 1985. "Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum." *Mind* 94 (373): 76-100. Reimpreso en S.P. Stich y T.A. Warfield, eds. 1994a.
——. 1986. "Information and Association," en M. Brand y R.M. Harnish, eds. 1986.
——. 1987. *Psychosemantics*. Cambridge, MA.: MIT Press.
——. 1989. "Making Mind Matter More." *Philosophical Topics* 17: 59-79.
——. 1990. *A Theory of Content and Other Essays*. Cambridge, MA.: MIT Press.

——. 1991. "You Can Fool Some of the People all of the Time, Everything Else Being Equal: Hedged Laws and Psychological Explanations." *Mind* 100 (1/397): 19-34.
——. 1994a. *The Elm and the Expert*. Cambridge, MA.: MIT Press.
——. 1994b. "Fodor, Jerry A.," en S. Guttenplan, ed. 1994a.
——. 1997. "Special Sciences: Still Autonomous After All These Years." *Noûs* 31 (*Mind, Causation, World. Philosophical Perspectives* 11): 149-163.
FODOR, J.A. y LEPORE, E. 1991. "Why Meaning (Probably) Isn't Conceptual Role." *Mind and Language* 6(4): 328-343. Reimpreso en S.P. Stich y T.A. Warfield, eds. 1994a.
FOSS, B.M., ed. 1966. *New Horizons in Psychology*. Harmondsworth: Penguin.
FOSTER, F. y SWANSON, J.W., eds. 1970. *Experience and Theory*. Londres: Duckworth.
FRANKENA, W.K. 1958. "Obligation and Motivation in Recent Moral Philosophy", en A.I. Melden, ed. 1958.
FREED, B. *et al*., eds. 1975. *Forms of Representation*. Amsterdam: North Holland / Elsevier.
FREGE, G. 1884. *Fundamentos de la Aritmética*. Traducción de U. Moulines. Barcelona: Laia, 1973. Reimpreso en G. Frege. *Frege: escritos filosóficos*.
——. 1891. "Función y concepto," en G. Frege. *Ensayos de semántica y filosofía de la lógica*.
——. 1892. "Sobre sentido y referencia," en G. Frege. *Ensayos de semántica y filosofía de la lógica*.
——. 1903. "Über die Grundlagen der Geometrie I/II." *Jahresbericht der Deutschen Mathematiker Vereinigung* 12: 319-324 / 368-375.
——. 1918. "El pensamiento: una investigación lógica," en G. Frege. *Ensayos de semántica y filosofía de la lógica*.
——. 1919. "La negación: una investigación lógica," en G. Frege. *Ensayos de semántica y filosofía de la lógica*.
——. *Wissenschaftlicher Briefwechsel*. Edición de G. Gabriel, H. Hermes, F. Kambartel, C. Thiel y A. Veraart. Hamburgo: Verlag Felix Meiner, 1976.
——. *Frege: escritos filosóficos*. Edición de U. Moulines. Barcelona: Crítica, 1996.
——. *Ensayos de semántica y filosofía de la lógica*. Edición de L.M. Valdés Villanueva. Madrid: Tecnos, 1998.
FRENCH, P.A., UEHLING, T.E., y WETTSTEIN, H.K., eds. 1977. *Studies in the Philosophy of Language (Midwest Studies in Philosophy II)*. Minneapolis, MN.: University of Minnesota Press.
——. 1979. *Contemporary Perspectives in the Philosophy of Language (Midwest Studies in Philosophy IV).* Minneapolis, MN.: University of Minnesota Press.
——. 1980. *Studies in Epistemology (Midwest Studies in Philosophy V).* Minneapolis, MN.: University of Minnesota Press.
——. 1984. *Causation and Causal Theories (Midwest Studies in Philosophy IX)*. Minneapolis, MN.: University of Minnesota Press.
——. 1986. *Studies in the Philosophy of Mind (Midwest Studies in Philosophy X)*. Minneapolis, MN.: University of Minnesota Press.
——. 1992. *The Wittgenstein Legacy (Midwest Studies in Philosophy XVII)*. Notre Dame, IN.: University of Notre Dame Press.
FREUD, S. *Obras completas*. Traducción de L. López-Ballesteros. Madrid: Biblioteca Nueva.
——.1919. *Lo siniestro*, en S. Freud, *Obras Completas*, VII.
FRIJDA, N.H. 1967. "Problems of Computer Simulation." *Behavioral Science* 12 (1): 59-66.
FRITSCH, G. y HITZIG, E. 1870. "Über die elektrische Erregbarkeit des Grosshirns." *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin* 37: 300-332.
GABOR, D. 1956. "Models in Cybernetics," en Accademia Nazionale dei Lincei, eds., 1956.
GALLISTEL, C.R. 1997. "Symbolic Processes in the Brain: the Case of Insect Navigation," en D. Scarborough, D.N. Osherson y S. Sternberg, eds. 1997.
GARCÍA CARPINTERO, M. 1995. "El funcionalismo," en F. Broncano, ed. 1995.
GARDNER, H. 1985. *The Mind's New Science: A History of the Cognitive Revolution*. Nueva York, NY.: Basic Books. Traducción de L. Wolfson, supervisada por A. Duarte: *La nueva ciencia de la mente: historia de la revolución cognitiva*. Barcelona: Paidós, 1987.
GALAVOTTI, M.C., ed., 2006. *Cambridge and Vienna: Frank P. Ramsey and the Vienna Circle*, 67–90. Dordrecht: Springer.

GEACH, P. 1957. *Mental Acts: their Contents and their Objects*. Londres: Routledge & Kegan Paul.
——. 1968. *A History of the Corruptions of Logic: Inaugural Lecture, January 22, 1968*. Leeds: Leeds University Press. Reimpreso en P. Geach 1972.
——. 1972. *Logic Matters*. Oxford: Blackwell.
GEIRSSON, H. y LOSONSKY, M., eds. 1996. *Readings in Language and Mind*. Oxford: Blackwell.
GERHARDT, H.C. 1983. "Communication and Environment," en T.R. Halliday y P.J.B. Slater, eds. 1983.
GETTIER, E.L. 1963. "Is Justified True Belief Knowledge?," *Analysis* 23: 121-123.
GIBSON, J.J. 1950. *The Perception of the Visual World*. Boston: Houghton Mifflin.
——. 1966. *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.
——. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.
GILLETT, C. 2004. "The Metaphysics of Realization, Multiple Realizability, and the Special Sciences." *Journal of Philosophy* 100: 591-603.
GÖDEL, K. 1934. "On Undecidable Propositions of Formal Mathematical Systems," apuntes de S.C. Kleene y J.B. Rosser editados en M. Davis, ed. 1965.
GODFREY-SMITH, P. 1994. "A Continuum of Semantic Optimism," en S.P. Stich y T.A. Warfield, eds. 1994a.
von GOETHE, J.W. 1808/1832. *Fausto*. Edición y traducción de R. Cansinos-Assens. Madrid: Aguilar, 1988.
GOLD, E. 1967. "Language Identification in the Limit." *Information and Control* 10: 447-474.
GOLDMAN, A.I. 1970. *A Theory of Human Action*. Englewood Cliffs, NJ.: Prentice-Hall.
——, ed. 1993. *Readings in Philosophy and Cognitive Science*. Cambridge, MA.: MIT Press.
GONDRA, J.M. 1992. "La génesis del modelo conductista," en J.A. Mora, ed. 1992a.
GONZÁLEZ, C. 2000. "Comprensión," en J. Muñoz y J. Velarde, eds. 2000.
GOODMAN, N. 1953. *Fact, Fiction, and Forecast*. Cambridge, MA.: Harvard University Press.
GOULD, S.J. y LEWONTIN, R.C. 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme." *Proceedings of the Royal Society of London: Biological Sciences* 205 (1161): 581-598.
GOULD, S.J. y VRBA, E.S. 1982. "Exaptation: A Missing Term in the Science of Form." *Paleobiology* 8: 4-15.
GRAHAM, G. 2007. "Behaviorism," en E.N. Zalta, ed. 2008.
GREGORY, R.L., ed. 1987. *The Oxford Companion to the Mind*. Oxford: Oxford University Press.
GREY WALTER, W.G. 1950. "An Imitation of Life." *Scientific American* 182 (5): 42-45.
——. 1951. "A Machine that Learns." *Scientific American* 184 (8): 60-63.
——. 1957. "Thinking Machines." *Institution of Production Engineers Journal* 36 (1): 4-12.
GRICE, P. 1957, "Meaning." *Philosophical Review* 66 (3): 377-388.
——. 1959. "Utterer's Meaning and Intentions". *Philosophical Review* 78 (2): 147-177.
GUERRERO DEL AMO, J.A. 2001a. "El Naturalismo Biológico de Searle," en P. Chacón *et al.* 2001.
——. 2001b. "Perspectivas actuales sobre la conciencia," en P. Chacón *et al.* 2001.
——. 2001c. "Problemas epistemológicos subyacentes a la teoría de la mente de Searle." *Logos: Anales del Seminario de Metafísica* 34: 297-316.
GUIJARRO, V. y GONZÁLEZ, L. 2010. *La quimera del autómata matemático. Del calculador medieval a la máquina analítica de Babbage*. Madrid: Cátedra.
van GULICK, R. 1989. "Metaphysical Arguments for Internalism and Why They Don't Work." En S. Silvers, ed. 1989.
GUNDERSON, K. 1971. *Mentality and Machines*. Garden City, NY.: Anchor.
——, ed. 1975. *Language, Mind, and Knowledge (Minnesota Studies in the Philosophy of Science VII)*. Minneapolis, MN.: University of Minnesota Press.
GUTHRIE, E.R. 1953. *The Psychology of Learning*. Nueva York, NY.: Harper.
——. 1959. "Association by Contiguity," en S. Koch, ed. 1959b.

GUTTENPLAN, S., ed. 1975. *Mind and Language (Wolfson College Lectures 1974)*. Oxford: Oxford University Press / Clarendon Press.
——, ed. 1994a. *A Companion to the Philosophy of Mind*. Cambridge, MA.: MIT Press.
——. 1994b. "Externalism / Internalism," en S. Guttenplan, ed. 1994a.
GÜZELDERE, G. 1997. "Approaching Consciousness," en N.J. Block, O.J. Flanagan y G. Güzeldere, eds. 1997.
HAECKEL, E. 1899. *Die Welträtsel. Gemeinverständliche Studien über monistische Philosophie*. Bonn: E. Strauss.
HABER, R.N., ed. 1968. *Contemporary Theory and Research in Visual Perception*. Nueva York, NY.: Holt, Rinehart & Winston.
HAHN, L. y SCHILPP, P., eds. 1986. *The Philosophy of W. v. O. Quine*. Chicago / La Salle, IL: Open Court.
HAHN, M. y STONESS, S.C., eds. 1999. *Proceedings of the 21st Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ.: Erlbaum.
HALDANE, J.J. 1987. "Reid, Scholasticism and Current Philosophy of Mind," en M. Dalgarno y E. Mathews, eds. 1987.
——. 1993. "Mind-World Identity and the Anti-Realist Challenge," en J.J. Haldane y C. Wright, eds. 1993.
HALDANE, J.J. y WRIGHT, C., eds. 1993. *Reality, Representation, and Projection*. Oxford: Oxford University Press / Clarendon Press.
HALL, R.J. 1996. "The Evolution of Color Vision without Colors." *Philosophy of Science* 63: 125-133.
HALLIDAY, T.R. y SLATER, P.J.B., eds. 1983. *Animal Behavior, vol.2: Communication*. Nueva York, NY.: Freeman.
HAMLYN, D.W. 1953. "Behaviour." *Philosophy* 28: 138-139.
HAMPSHIRE, S. 1950. "Critical Notice of Ryle, *The Concept of Mind*." *Mind* 59 (234): 237-255.
HANSON, N.R. 1958. *Patterns of Discovery*. Cambridge: Cambridge University Press.
HARLOW, H.F. 1949. "The Formation of Learning Sets." *Psychological Review* 56: 5-65.
——. 1953a. "Motivation as a Factor in the Acquisition of New Responses", en J.S. Brown *et al*., eds. 1953.
——. 1953b. "Mice, Monkeys, Men, and Motives." *Psychological Review* 60: 23-32.
——. 1958. "The Nature of Love." *American Psychologist* 13: 673-685.
HARMAN, G. 1973. *Thought*. Princeton, NJ.: Princeton University Press.
——. 1974. "Meaning and Semantics," en M.R. Munitz y P.K. Unger, eds. 1974.
——. 1977. *Thought*. Princeton, NJ.: Princeton University Press.
——. 1982. "Conceptual Role Semantics." *Nôtre Dame Journal of Formal Logic* 23: 242-256.
——. 1987. "(Non-Solipsistic) Conceptual Role Semantics," en E. LePore, ed. 1987.
HARRISON, J. 1963. "Does Knowing Imply Believing?" *Philosophical Quarterly* 13 (53): 322-332.
HATFIELD, G. 1989. "Computation, Representation, and Content in Noncognitive Theories of Perception," en S. Silvers, ed. 1989.
HAUGELAND, J. 1978. "The Nature and Plausibility of Cognitivism." *Behavioral and Brain Sciences* 2: 215-226. Reimpreso en J. Haugeland 1998.
——, ed. 1981. *Mind Design*. Cambridge, MA.: MIT Press.
——. 1982. "The Mother of Intention." *Noûs* 16: 613-619.
——. 1998. *Having Thought. Essays on the Metaphysics of Mind*. Cambridge, MA.: Harvard University Press.
——. 2002a. "Andy Clark on Cognition and Representation," en H. Clapin, ed. 2002a.
——. 2002b. "Reply to Cummins on Representation and Intentionality," en H. Clapin, ed. 2002a.
HAYES, J.R. y SIMON, H.A. 1976. "The Understanding Process: Problem Isomorphs." *Cognitive Psychology* 8: 165-180.
HAZEWINKEL, M., ed. 1987-2002. *Encylopaedia of Mathematics*. Berlín: Springer-Verlag.
HEATH, A.F., ed. 1981. *Scientific Explanations*. Oxford: Oxford University Press / Clarendon Press.

HEBB, D.O. 1960. "The American Revolution." *American Psychologist* 15: 735-745.
——. 1968. *Psicología.* México, D.F.: Interamericana.
HEIDEGGER, M. 1927. *El ser y el tiempo*. Traducción de J. Gaos. México, D.F.: Fondo de Cultura Económica, 1951.
van HEIJENOORT, J. 1967. *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931*. Cambridge, MA.: Harvard University Press. Tercera edición corregida, 1976.
HEIL, J. 1983. *Perception and Cognition*. Berkeley, CA.: University of California Press.
——, ed. 1989a. *Cause, Mind, and Reality*. Dordrecht: Kluwer.
——. 1989b. "Intentionality Speaks for Itself," en S. Silvers, ed. 1989.
HEIL, J. y MELE, A., eds. 1993. *Mental Causation*. Oxford: Oxford University Press / Clarendon Press.
HEMPEL, C.G. 1935. "Analyse logique de la psychologie." *Revue de Synthèse* 10: 27-42. Reimpreso como "The Logical Analysis of Psychology" en N.J. Block, ed. 1980. Traducción de W. Sellars.
——. 1942. "The Function of General Laws in History." *Journal of Philosophy* 39: 35-48.
——. 1958. "The Theoretician's Dilemma," en H. Feigl, M. Scriven y G. Maxwell, eds. 1958.
——. 1965. *Aspects of Scientific Explanation*. Nueva York, NY.: Free Press.
HERBRAND, J. 1932. "Sur la non-contradiction de l'arithmetique." *Journal fur die reine und angewandte Mathematik* 166 : 1-8.
HERMOSO, J. 2001a. "La teoría representacional de la mente de Jerry Fodor," en P. Chacón *et al*. 2001.
——. 2001b. "Una mente en movimiento: reseña de *Tipos de Mentes*, de D.C. Dennett." *Anábasis* 2-3: 137-146.
——. 2005. "Ángeles y autómatas: Descartes y los límites de la imaginación," en M. Rodríguez, ed. 2005.
HERMOSO, J. y CHACÓN, P. 2000. "La irreductibilidad de la intencionalidad y los dos conceptos del Trasfondo en Searle," en P. Chacón y M. Rodríguez, eds. 2000.
HIERRO-PESCADOR, J. 2002. *Filosofía de la mente y de la Ciencia Cognitiva*. Madrid: Akal.
HILBERT, D. 1901. "Mathematische Probleme," A*rchiv der Mathematik und Physik* 3 (1): 44-63, 213–237, reimpreso en D. Hilbert, *Gesammelte Abhandlungen. III*. Berlín: Springer Verlag, 1935. Traducción al inglés de M. Winton (1902) en *Bulletin of the American Mathematical Society* 8: 437–479. URL = <http://aleph0.clarku.edu/~djoyce/hilbert/problems.html, 29 de agosto de 2010.
HILBERT, D. y ACKERMANN, W. 1928. *Grundzüge der Theoretischen Logik*. Berlín: Springer Verlag.
HILGARD, E.R. y BOWER, G.H. 1966. *Teorías del aprendizaje*. Traducción de J.M. Salazar Palacios. México, D.F.: Trillas, 1976.
HISPANO, P. *Tractatus, llamados después Summulæ logicales*. Traducción de M. Beuchot. México D.F.: UNAM, 1986. Basado en *Tractatus, called afterwards Summulæ logicales*. Traducción, edición crítica e introducción de L.M. de Rijk. Assen: Van Gorcum, 1972.
HODGES, A., 1983. *Alan Turing: the Enigma*. Londres: Burnett.
HOCHBERG, J. 1968. "In the Mind's Eye," en R.N. Haber, ed. 1968.
HOFFMANN, E.T.A. 1816. "El hombre de la arena", en *El hombre de la arena: trece historias siniestras y nocturnas*. Traducción de L.F. Moreno. Madrid: Valdemar, 1998. Reimpreso en S. Bueno y M. Peirano, eds. 2009.
HOFSTATDER, D.R. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid*. Nueva York, NY.: Basic Books.
HOLT, E.B. 1914. *The Concept of Consciousness*. Nueva York, NY.: Macmillan.
HOMERO. *Ilíada*. Introducción de J. de Hoz y traducción de L. Segalá. Madrid: Espasa Calpe, 1954.
HONDERICH, T., ed. 1973. *Essays on Freedom of Action*. Londres: Routledge & Kegan Paul.
HOOK, S., ed. 1960. *Dimensions of Mind*. Nueva York, NY.: New York University Press.
HOOKER, C. 1981. "Towards a General Theory of Reduction, III: Cross-Categorial Reductions." *Dialogue* 20: 496-529.

HOOKWAY, C., ed. 1984. *Minds, Machines, and Evolution*. Cambridge: Cambridge University Press.
HOPKINS, C.D. 1983. "Sensory Mechanisms in Animal Communication," en T.R. Halliday y P.J.B. Slater, eds. 1983.
HORGAN, T. 1993. "Nonreductive Materialism and the Explanatory Autonomy of Psychology," en S. Wagner y R. Warner, eds. 1993.
——. 1994. "Computation and Mental Representation," en S.P. Stich y T.A. Warfield, eds. 1994a.
HORST, S.W. 1996. *Symbols, Computation and Intentionality. A Critique of the Computational Theory of Mind.* Berkeley: University of California Press.
HORWICH, P. 1984. "Critical notice: Saul Kripke: *Wittgenstein on Rules and Private Language*." *Philosophy of Science* 51: 1.
HUARTE DE SAN JUAN, J. 1575 / 1590. *Examen de ingenios para las ciencias*. Edición, introducción y notas de F. Fresco. Madrid: Espasa, 1991.
HULL, C.L. 1930. "Knowledge and Purpose as Habit Mechanisms." *Psychological Review* 37: 511.525.
——. 1935. "The Mechanism of the Assembly of Behavior Segments in Novel Combination Suitable for Problem Solution." *Psychological Review* 42: 219-245.
——. 1943. *Principles of Behavior*. Nueva York, NY.: Appleton-Century-Crofts.
HULL, C.L. y BAERSTEIN, H.D. 1929. "A Mechanical Parallel to the Conditioned Reflex." *Science* 70: 14-15.
HUME, D. 1739. *A Treatise of Human Nature*. Edición de P.H. Nidditch, sobre la anterior de L.A. Selby-Biggs. Oxford / Nueva York, NY.: Oxford University Press, 1978.
HUXLEY, L. 1900. *The Life and Letters of Thomas Henry Huxley*. Nueva York, NY.: Appleton.
HUXLEY, T.H. 1866. *Lessons in Elementary Physiology*. Londres: Macmillan. Tercera edición: 1872.
——. 1870. "Has a Frog a Soul, and of What Nature is That Soul, Supposing It to Exist?" *Metaphysical Society Papers* (inéditos). Bodleian Library, Oxford, 2657e.1. URL = <http://aleph0.clarku.edu/huxley/Mss/FROG.html, 8 de febrero de 2010.
——. 1876. "On the Hypothesis that Animals are Automata." *Fortnightly Review* 22: 558-580. Reimpreso en T.H. Huxley 1893.
——. 1893. *Collected Essays*. Londres: Macmillan.
HYMAN, A., ed. 1989. *Science and Reform: Selected Works of Charles Babbage*. Cambridge: Cambridge University Press.
IBARRA, A. 2000. "La naturaleza vicarial de las representaciones," en A. Ibarra y T. Mormann, ed. 2000.
IBARRA, A. y MORMANN, T., eds. 2000. *Variedades de la representación en ciencia y filosofía*. Barcelona: Ariel.
JACKENDOFF, R. 1987. *Consciousness and the Computational Mind*. Cambridge, MA.: MIT Press.
JACKSON, F. 1982. "Epiphenomenal Qualia." *Philosophical Quarterly* 32: 127-136.
——. 1986. "What Mary Didn't Know." *Journal of Philosophy* 83: 291-295.
——. 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press.
JACKSON, F. y PETTIT, P. 1990. "Causation in the Philosophy of Mind." *Philosophy and Phenomenological Research* 50: 195-214.
JACOB, P. 1987. "Thoughts and Beliefs Ascriptions." *Mind and Language* 2 (4): 301-325. Reimpreso en S. Silvers, ed. 1989.
——. 2002. "Some Problems for Reductive Physicalism." *Philosophy and Phenomenological Research* 65: 648-654.
JAMES, W. 1877. "Review of *The Functions of the Brain*, by David Ferrier; *The Physiology of Mind*, by Henry Maudsley; and *Le Cerveau et ses fonctions*, by Jules Luys". Reimpreso en W. James, *Essays, Comments, and Reviews*.
——. 1884. "¿Qué es una emoción?" Traducción de E. Gaviria. *Estudios de Psicología* 21: 57-73 (1983).

——. 1890. *The Principles of Psychology*. Nueva York, NY.: Holt.
——. 1892. *Psychology: the Briefer Course*. Nueva York, NY.: Holt.
——. 1907. *El pragmatismo: un nuevo nombre para viejas formas de pensar*. Traducción de R. del Castillo. Madrid: Alianza, 2007.
——. 1909. *The Meaning of Truth: A Sequel to 'Pragmatism'*. Nueva York: Longmans, Green, & Co.
——. 1912. *Essays in Radical Empiricism*. Nueva York, NY.: Longmans, Green and Co. Reimpreso con correcciones en Mineola, NY.: Dover, 2003.
——. *Pragmatism and the Meaning of Truth*. Cambridge, MA.: Harvard University Press, 1975.
——. *Essays, Comments, and Reviews*. Cambridge, MA.: Harvard University Press, 1987.
JEFFRESS, L.A., ed. 1951. *Cerebral Mechanisms in Behavior*. Nueva York, NY.: Wiley.
JENKINS, J.J. 1968. "The Challenge to Psychological Theorists," en T.R. Dixon y D.C. Horton, eds. 1968.
——. 1986. "Interview with James J. Jenkins," en B.J. Baars 1986.
JENNINGS, H.S. 1904. "Physical Imitations of the Activities of *Amoeba*." *American Naturalist* 38: 625-642.
——. 1906. *The Behavior of the Lower Organisms*. Nueva York, NY.: Columbia University Press. Segunda edición con introducción de D.D. Jensen: Bloomington, IN.: University of Indiana Press, 1962.
——. 1910. "Diverse Ideas and Divergent Conclusions in the Study of Behavior in Lower Organisms." *American Journal of Psychology* 21(3): 349-370.
JENTSCH, E. 1906. "Zur Psychologie des Unheimlichen." *Psychiatrisch-Neurologische Wochenschrift* 8(22): 195-198 / 8(23): 203-205.
JOHNSON, W.E. 1921/1924. *Logic*. Cambridge: Cambridge University Press.
JOHNSON-LAIRD, P. 1983. *Mental Models: Toward a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA.: Harvard University Press.
——. 1987. "How Could Consciousness Arise from the Computations of the Brain?" en C. Blakemore y S. Greenfield, S., eds.1987.
——. 1988. *El ordenador y la mente. Introducción a la ciencia cognitiva*. Traducción de A. Medina. Barcelona: Paidós, 1990.
JOHNSON-LAIRD, P.N. y STEEDMAN, M.J. 1978. "The Psychology of Syllogisms." *Cognitive Psychology* 10: 64-99.
KAHNEMAN, D. y TVERSKY, A. 1973. "On the Psychology of Prediction." *Psychological Review* 80: 273-251.
——. 1984. "Choices, Values, and Frames." *American Psychologist* 34(4): 341-350.
KAMIN, L.J. 1969. "Selective Association and Conditioning," en N.J. McKintosh y W.K. Honig, eds. 1969.
KANT, I. 1781. *Crítica de la Razón Pura*. Traducción de P. Ribas. Madrid: Alfaguara, 1978/2003.
KAPLAN, D. 1968. "Quantifying in." *Synthese* 19: 178-214.
——. 1979a. "Dthat," en P. Cole, ed. 1979.
——. 1979b. "On the Logic of Demonstratives." *Journal of Philosophical Logic* 8: 81-98. Reimpreso en P.A. French, T.E. Uehling, y H.K. Wettstein, eds. 1979.
——. 1989a. "Demonstratives," en J. Almog, J. Perry y H.K. Wettstein, eds. 1989.
——. 1989b. "Afterthoughts," en J. Almog, J. Perry y H.K. Wettstein, eds. 1989.
KEMENY, J.G. y OPPENHEIM, P. 1956. "On Reduction." *Philosophical Studies* 7: 6-19.
KENDLER, T.S. y KENDLER, H.H. 1962. "Vertical and Horizontal Processes in Problem Solving." *Psychological Review* 69: 1-16.
KENNY, A. 1992. *The Metaphysics of Mind*. Oxford: Oxford University Press / Clarendon Press.
——. 1993. *Tomás de Aquino y la mente*. Traducción de J.M. López de Castro. Barcelona: Herder, 2000.
——. 1995. *Introducción a Frege*. Traducción de C. García Trevijano. Madrid: Cátedra, 1997.
KEPLER, J. 1604. *Ad Vitellionem paralipomena, quibus astronomiae pars optica traditur*. Frankfurt: Claudium Marnium & Haeredes Ioannis Aubrii.

KHAS'MINSKII, R.Z. 2001. "Markov Process," en Hazewinkel, M., ed. 1987-2002.
KIM, J. 1972. "Phenomenal Properties, Psychophysical Laws, and the Identity Theory." *The Monist* 56: 177-192. Reimpreso parcialmente en N.J. Block, ed. 1980 con el título "Physicalism and the Multiple Realizability of Mental States."
——. 1977. "Perception and Reference without Causality." *Journal of Philosophy* 74: 606-620.
——. 1978. "Supervenience and Nomological Inconmmensurables." *American Philosophical Quarterly* 15 (2): 149-156.
——. 1979. "Causality, Identity, and Supervenience in the Mind-Body Problem," en P.A. French, T.E. Uehling y H.K. Wettstein, eds. 1979.
——. 1982. "Psychological Supervenience as a Mind-Body Theory." *Cognition and Brain Theory* 5: 129-147.
——. 1984a. "Concepts of Supervenience." *Philosophy and Phenomenological Research* 45(2): 153-176.
——. 1984b. "Epiphenomenal and Supervenient Causation," en P.A. French, T.E. Uehling y H.K. Wettstein, eds. 1984.
——. 1988. "What is Naturalized Epistemology?" *Philosophical Perspectives* 2: 381-406.
——. 1989. "The Myth of Non-Reductive Materialism." *Proceedings and Addresses of the American Philosophical Association* 63: 31-47. Reimpreso en J. Kim, 1993a.
——. 1992. "Multiple Realization and the Metaphysics of Reduction." *Philosophy and Phenomenological Research* 52: 1-26. Reimpreso en J. Kim, 1993a [citado según la edición de 1992].
——. 1993a. *Supervenience and Mind*. Cambridge: Cambridge University Press.
——. 1993b. "Can Supervenience and 'Non-Strict Laws' Save Anomalous Monism?", en J. Heil y A. Mele, eds. 1993.
——. 1993c. "The Non-Reductivist Troubles with Mental Causation," en J. Heil y A. Mele, eds. 1993.
——. 1998. *Mind in a Physical World*. Cambridge, MA.: MIT Press.
——. 2002. "*Précis of* Mind in a Physical World." *Philosophy and Phenomenological Research* 65: 640-643.
KIMBLE, G.A. 1998. "Introduction to the Transaction Edition," en J.B. Watson, 1924/1930.
KINTSCH, W., MILLER, J.R. y POLSON, R.E. 1984. *Method and Tactics in Cognitive Science*. Hillsdale, NJ.: Lawrence Earlbaum.
KIRK, G.S., RAVEN, J.E. y SCHOFIELD, M., eds. (1983). *Los filósofos presocráticos: historia crítica con selección de textos*. Traducción de J. García Fernández. Madrid : Gredos, 1987.
KLAVANS, J. y RESNIK, P., eds. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* Cambridge, MA.: MIT Press.
KLEENE, S.C. 1935. "A Theory of Positive Integers in Formal Logic," *American Journal of Mathematics* 57: 153-173, 219-244.
——. 1936. "Lambda-Definability and Recursiveness," *Duke Mathematical Journal* 2: 340-353.
KOCH, S., ed. 1959a. *Psychology: a Study of a Science, vol. I*: *Conceptual and Systematic Study. Sensory, Perceptual, and Physiological Formulations*. Nueva York, NY.: McGraw-Hill.
——. 1959b. *Psychology: a Study of a Science, vol. II: Conceptual and Systematic Study. Systematic Formulation, Learning, and Special Processes*. Nueva York, NY.: McGraw-Hill.
——. 1959c. *Psychology: a Study of a Science, vol. III*: *Conceptual and Systematic Study. Formulations of the Person and the Social Context*. Nueva York, NY.: McGraw-Hill.
——. 1959d. "Epilogue: Some Trends in Study" en S. Koch, ed. 1959c.
KÖHLER, W. 1917. *Experimentos sobre la inteligencia de los chimpancés*. Traducción de J.C. Gómez. Madrid: Debate, 1989.
KRUEGER , R.G. y HULL, C.L. 1931. "An Electrochemical Parallel to the Conditioned Reflex." *Journal of General Psychology* 5: 262-269.
KRIPKE, S.A. 1972/1980. *El nombrar y la necesidad*. Traducción de M.M. Valdés. México, D.F.: UNAM, 1995.
——. 1982. *Wittgenstein: a propósito de reglas y lenguaje privado. Una exposición elemental*. Traducción de J. Rodríguez Marqueze. Madrid: Tecnos, 2006.

KUENNE, M. 1946. "Experimental Investigation of the Relation of Language to Transposition Behavior in Young Children." *Journal of Experimental Psychology* 36: 471-490.
KUHN, T.S. 1962. *La estructura de las revoluciones científicas*. Traducción de A. Contin. México D.F.: Fondo de Cultura Económica, 1971. Incluye T.S. Kuhn, 1969. "Posdata."
LACHMAN, T., LACHMAN, J.L. y BUTTERFIELD, E.C. 1979. *Cognitive Psychology and Information Processing: an Introduction*. Hillsdale, NJ.: Erlbaum.
LAKATOS, I. 1978. *La metodología de los programas de investigación científica. (Escritos filosóficos 1)*. Edición de J. Woray y G. Currie; traducción de J.C. Zapatero. Madrid: Alianza, 2007.
LAKOFF, G. y JOHNSON, M. 1980. *Metáforas de la vida cotidiana*. Traducción de C. González. Madrid: Cátedra, 1986.
LAMIELL, J.T. 1998. "'Nomothetic' and 'Idiographic': Contrasting Windelband's Understanding with Contemporary Usage." *Theory and Psychology* 8 (1): 23-28.
LANDAUER, T.K., ed. 1967. *Readings in Physiological Psychology: The Bodily Basis of Behavior*. Nueva York, NY.: McGraw-Hill.
LANGACKER, R.W. 1987. *Foundations of Cognitive Grammar, I: Theoretical Prerequisites*. Palo Alto, CA.: Stanford University Press.
LASHLEY, K.S. 1917. "The Accuracy of Movement in the Absence of Excitation from the Moving Organ." *American Journal of Physiology* 43: 169-194.
——. 1929. *Brain Mechanisms and Intelligence*. Chicago, IL.: University of Chicago Press.
——. 1930. "Basic Neural Mechanisms in Behavior." *Psychological Review* 37: 1-24.
——. 1937. "Functional Determinants of Cerebral Localization." *Archives of Neurology and Psychiatry* 38: 371-387,
——. 1942. "The Problem of Cerebral Organization in Vision." *Biological Symposia* 7: 301-322.
——. 1951. "The Problem of Serial Order in Behavior," en L.A. Jeffress, ed. 1951.
LEAHEY, T.H. 2005. *Historia de la psicología. Principales corrientes del pensamiento psicológico*. Traducción de M. de Ancos y C. Rivera. Madrid: Pearson / Prentice Hall. Sexta edición.
LEBLANC, H. ed. 1973. *Truth, Syntax, and Modality*. Amsterdam: North Holland / Elsevier.
LEHRER, K., ed. 1966. *Freedom and Determinism*. Nueva York, NY.: Random House.
LEIBNIZ, G.W. 1666. *Dissertatio de Arte Combinatoria (On the Art of Combination)*, en G.W. LEIBNIZ. *Leibniz: Logical Papers*. Edición de G.H.R. Parkinson. Oxford: Oxford University Press, 1966.
——. 1678. *Elementa Characteristicae Universalis*, en G.W. LEIBNIZ. *Opuscules et fragments inédites de Leibniz*. Edición de L. Couturat (1903). Hildesheim: Georg Olms, 1961.
——. 1715. *Monadologie (Monadología)*, en G.W. LEIBNIZ. *Escritos filosóficos*. Edición de E. de Olaso. Madrid: Antonio Machado, 2003.
LEOPOLD, D.A. 2009. "Neuroscience: Pre-emptive Blood Flow." *Nature* 457: 387-388.
LEPORE, E., ed. 1987. *New Directions in Semantics*. Londres: Academic Press.
LEPORE, E. y LOEWER, B. 1987. "Dual Aspect Semantics," en E. LePore, ed. 1987. Reimpreso en S.Silvers, ed. 1989.
——. "More on Making Mind Matter." *Philosophical Topics* 17: 175-191.
LEVIN, J. 2004. "Functionalism," en E.N. Zalta, ed. 2008.
LEWIN, K. 1951. "Intention, Will, and Need," en D. Rapaport, ed. 1951.
LEWIS, C.I. 1912. "Implication and the Algebra of Logic." *Mind* 21(84): 522-531.
——. 1929. *Mind and the World-Order: Outline of a Theory of Knowledge*. Nueva York, NY.: Charles Scribsner. Segunda edición: 1956. Nueva York, NY.: Dover.
LEWIS, D.K. 1966. "An Argument for the Identity Theory." *Journal of Philosophy* 63: 17-25. Traducción castellana: "Un argumento a favor de la teoría de identidad." *Cuadernos de Crítica* 30.
——. 1969. "Review of *Art, Mind, and Religion* [W.H Capitan y D.D Merrill, eds. 1967]" *Journal of Philosophy* 66: 23-25. Reimpreso parcialmente en N.J. Block, ed. 1980.
——. 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy* 50 (3): 248-259. Reimpreso en N.J. Block, ed. 1980.

——. 1980. "Mad Pain and Martian Pain," en N.J. Block, ed. 1980.
——. 1983a. *Philosophical Papers, vol. 1*. Oxford: Oxford University Press.
——. 1983b. "Postscript to 'Mad Pain and Martian Pain'," en D.K. Lewis, 1983a.
LEWONTIN, R.C. 1978. "Adaptation." *Scientific American* 239: 212-230.
LIBET, B., GLEASON, C. A., WRIGHT, E. W., y PEARL, D. K. 1983. "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). The Unconscious Initiation of a Freely Voluntary Act." *Brain* 106: 623-642.
LIBET, B. 1985. "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *Behavioral and Brain Sciences* 8: 529-566.
LIGHTHILL, J. 1973. "Artificial Intelligence: A General Survey," en *Artificial Intelligence: a paper symposium*, Londres: Science Research Council.
LINDBERG, D. 1976. *Theories of Vision from Al-Kindi to Kepler*. Chicago, IL.: Chicago University Press.
LIZ, M. 1995. "Causalidad y contenido mental," en F. Broncano, ed. 1995.
——., ed. 2012. *Puntos de vista: una investigación filosófica*. Barcelona: Laertes.
LLANO, A. 1999. *El enigma de la representación*. Madrid: Síntesis.
LLEDÓ, E. 2001. *El origen del diálogo y la ética. Una introducción al pensamiento de Platón y Aristóteles*. Madrid: Gredos.
LLOYD MORGAN, C. 1896. *Habit and Instict*. Londres: Arnold.
——. 1900. *Animal Behaviour*. Londres: Arnold.
LOBACHEVKSI, N.I. 1832. "O nachalah geometrii." *Kazanski Vestnik* 25, 27-28.
LOEB, J. 1918. *Forced Movements, Tropisms, and Animal Conduct*. Philadelphia: Lippincott.
LOCKE, J. 1689. *Ensayo sobre el entendimiento humano*. Traducción de Edmundo O'Gorman. México, D.F.: Fondo de Cultura Económica, 1999.
LOEWER, B. y REY, G., eds. 1991. *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.
LONGUET-HIGGINS, H.C. 1973. "Comments on the Lighthill Report and the Sutherland Reply," en *Artificial Intelligence: a paper symposium*, Londres: Science Research Council.
LÓPEZ DE LA VIEJA, T. 2009. "Comprensión," en R. Reyes, ed. 2009.
LÓPEZ-ORNAT, S. y GALLO, P. 2004. "Acquisition, Learning, or Development of Language? Skinner's *Verbal Behavior* Revisited." *Spanish Journal of Psychology*: 7 (2): 161-170.
LOTKA, A.J. 1925. *Elements of Physical Biology*. Baltimore: William and Wilkins.
LUCE, R.D., BUSH, R.R. y GALANTER, E., eds. 1963. *Handbook of Mathematical Psychology, vol. 1*. Nueva York, NY.: Wiley.
LUCRECIO CARO, T. *De la naturaleza de las cosas*. Traducción de José Marchena (Abate Marchena), introducción de Agustín García Calvo y notas de Domingo Plácido. Madrid: Cátedra, 1990.
LUDWIG, K. 1852-1856. *Lehrbuch der Physiologie des Menschen*. Leipzig / Heidelberg: C.F. Winter'sche Verlagshandlung.
LYCAN, W. 1981. "Form, Function, and Feel." *Journal of Philosophy* 78: 24-49.
——. 1986a. "Tacit Belief," en R.J. Bogdan, ed. 1986.
——. 1986b. "Thoughts about Things," en M. Brand y R.M. Harnish, eds. 1986.
——. 1987. *Consciousness*. Cambridge, MA.: MIT Press.
——. 1988. *Judgement and Justification*. Cambridge: Cambridge University Press.
——. 1993. "A Deductive Argument for the Representational Theory of Thinking," *Mind and Language* 8 (3): 404-420.
——. 1994. "Functionalism,"en S. Guttenplan, ed. 1994a.
MACDONALD, G. y WRIGHT, C., eds. 1987. *Fact, Science, and Morality: Essays on A.J. Ayer's Language, Truth, and Logic*. Oxford: Blackwell.
MACHLUP, F. y MANSFIELD, U., eds. 1984. The *Study of Information: Interdisciplinary Messages*. Nueva York: Wiley.
MACKAY, D.M. 1952. "Mentality in Machines." *Proceedings of the Aristotelian Society (Supplements)* 26: 61-86.
——. 1954. "On Comparing the Brain with Machines." *Advancement of Science* 10: 402-406.

——. 1962. "The Use of Behavioral Language to Refer to Mechanical Processes." *British Journal for the Philosophy of Science* 13: 89-103.
MACKENZIE, B.D. 1977. *Behaviorism and the Limits of the Scientific Method*. Londres: Routledge & Kegan Paul.
MACKIE, J.L. 1976. *Problems from Locke*. Oxford: Oxford University Press / Clarendon Press.
MALCOLM, N. 1959. *Dreaming*. Londres: Routledge & Kegan Paul.
——. 1971. "The Myth of Cognitive Processes and Structures," en T. Mischel, ed. 1971.
——. 1984. "Consciousness and Causality," en D.M. Armstrong y N. Malcolm, eds. 1984.
MANDLER, G. 2002. "Origins of the Cognitive (R)evolution." *Journal of History of the Behavioral Sciences* 38 (4): 339-353.
MARCEL, A. y BISIACH, M., eds. 1988. *Consciousness in Contemporary Science*. Oxford: Oxford University Press.
MARCO AURELIO. *Meditaciones*. Traducción de R. Bach Pellicer. Madrid: Gredos, 1994.
MARR, D. 1982. *La visión. Una investigación basada en el cálculo acerca de la representación y el procesamiento humano de la información visual.* Traducción de T. del Amo. Madrid: Alianza, 1985.
MARRAS, A. 2006. "Emergence and Reduction: Reply to Kim." *Synthese* 151: 561-569.
MARSHALL, A. 1890. *Principles of Economy*. Londres: Macmillan.
MARTEL JOHNSON, D. 1997. "What Is the Purported Discipline of Cognitive Science and Why Does it Need to Be Reassessed at the Present Moment? The Search for 'Cognitive Glue'," en D. Martel Johnson y C.E. Erneling, eds. 1997.
MARTEL JOHNSON, D. y ERNELING, C.E., eds. 1997. *The Future of the Cognitive Revolution*. Oxford: Oxford University Press.
MARTIN, C.B. 1994. "Dispositions and Conditionals." *Philosophical Quarterly* 44: 1-8.
MARTIN, M. 1973. "Are Cognitive Processes and Structures a Myth?" *Analysis* 33 (3): 83-88.
MARTÍNEZ-FREIRE, P.F. 1995. *La nueva filosofía de la mente*. Barcelona: Gedisa.
——. 2001. "Base empírica y teoría funcionalista en las ciencias cognitivas"; "Respuesta a Antonio Blanco Salgueiro." *Ágora. Papeles de Filosofía* 20(1): 87-104; 113-114.
MARX, K. y ENGELS, F. 1848. *Manifiesto del Partido Comunista*. Traducción de J. Muñoz. Milán: Mondadori / Silvio Berlusconi Editore, 1998.
MARX, M.H. 1951. *Psychological Theory: Contemporary Readings*. Nueva York, NY.: Macmillan.
MARX, M.H. y HILLIX, W.A. 1963. *Sistemas y teorías psicológicas contemporáneos*. Traducción de J. Colapinto, supervisada por E. Butelman, según la tercera edición revisada de 1979. Buenos Aires: Paidós, 1983.
MATTHEWS, R.J. 1989. "The Alleged Evidence for Representationalism," en S. Silvers, ed. 1989.
MAXWELL, G. 1962. "The Ontological Status of Theoretical Entities," en H. Feigl y G. Maxwell, eds. 1962.
MAYR, E. 1974. "Behavioral programs and evolutionary strategies." *American Scientist* 62: 650-659.
MEDLIN, B. 1967. "Ryle and the Mechanical Hypothesis," en C.F. Presley, ed. 1967.
von MELCHNER, L., PALLAS, S. y SUR, M. 2000. "Visual Behaviour Mediated by Auditory Cortex Redirected to the Auditory Pathway." *Nature* 404: 871-876.
MELTZER, B. y MICHIE, D. 1969a. *Machine Intelligence, IV*. Nueva York, NY.: Elsevier.
——. 1969b. *Machine Intelligence, V*. Edimburgo: Edinburgh University Press.
McCARTHY, J., MINSKY, M., ROCHESTER, N. y SHANNON, C.E. 1955. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." URL = < http://www-formal.stanford.edu/jmc/history/dartmouth.pdf, 13 abril de 2011.
McCARTHY, J. y HAYES, P. 1969. "Some Philosophical Problems from the Perspective of Artificial Intelligence," en B. Meltzer y D. Michie, eds. 1969a.
McCORQUODALE, K. 1970. "On Chomsky's Review of Skinner's *Verbal Behavior*." *Journal of the Experimental Analysis of Behavior* 13: 83-99.
McCULLOCH, W.S. y PITTS, W.H. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 7: 115-133.

McDERMOTT, D. 1976. "Artificial Intelligence Meets Natural Stupidity." *SIGART Newsletter* 57. Reimpreso en J. Haugeland, ed. 1981.
McDERMOTT, M. 1986. "Narrow Content." *Australasian Journal of Philosophy* 64 (3): 277-288. Reimpreso en S. Silvers, ed. 1989.
McDONALD, G.F. 1979. *Perception and Identity*. Londres: Macmillan.
McDOUGALL, W. 1911. *Body and Mind. A History and a Defence of Animism*. Londres: Methuen.
——. 1928. "Men or Robots?," en C. Murchison, ed. 1928.
McDOWELL, J. 1992. "Meaning and Intentionality in Wittgenstein's Later Philosophy," en P.A. French, T.E. Uehling y H.K. Wettstein, eds. 1992.
——. 1994a. *Mente y Mundo*. Traducción de M.A. Quintana. Salamanca: Sígueme, 2003.
——. 1994b. "The Content of Perceptual Experience." *Philosophical Quarterly* 44: 190-205.
McGINN, C. 1982. "The Structure of Content," en A. Woodfield, ed. 1982.
——. 1989. "Can We Solve the Mind-Body Problem?," *Mind* 98: 349-366.
——. 1999. *The Mysterious Flame: Conscious Minds in a Material World*. Nueva York, NY.: Basic Books.
McKAY, A. y MERRILL, D.D.. 1976. *Issues in the Philosophy of Language*. New Haven: Yale University Press.
McKINTOSH, N.J. y HONIG, W.K. eds. 1969. *Fundamental Issues in Associative Learning*. Halifax: Dalhousie University Press.
McLAUGHLIN, B., BECKERMANN, A. y WALTER, S., eds. 2009. *Oxford Handbook in the Philosophy of Mind*. Oxford: Oxford University Press.
McTAGGART, J. 1908. "The Unreality of Time." *Mind* 17: 457-474.
MEINONG, A. 1904. "Über Gegenstandstheorie," en A. Meinong *et al.* 1904.
MEINONG, A. *et al.* 1904. *Untersuchungen zur Gegenstandstheorie und Psychologie*. Leipzig: Verlag von Johannes Ambrosious Barth, 1904.
MELDEN, A.I., ed. 1958. *Essays in Moral Philosophy*. Seattle, WA.: University of Washington Press.
MERRILL, D.D. y GRIMM, R.H. eds.1988. *Contents of Thought. Proceedings of the 1985 Oberlin Colloquium in Philosophy*. Tucson, AZ.: University of Arizona Press.
de la METTRIE, J.O. 1748. *El hombre máquina*. Edición de J.L. Pérez. Madrid: Alhambra, 1987.
MEYER, M.F. 1908. "The Nervous Correlate of Attention." *Psychological Review* 15: 358-372.
——. 1911. *The Fundamental Laws of Human Behavior*. Boston: R.G. Badger. Reeditado con un estudio introductorio por R.H. Wozniak. Londres: Routledge & Kegan Paul, 1993.
——. 1912. "The Present Status of the Problem of the Relations between Mind and Body." *Journal of Philosophy* 9: 365-371.
——. 1913. "The Comparative Value of Various Conceptions of Nervous Function Based on Mechanical Analogies." *American Journal of Psychology* 24: 555-563.
——. 1921. *The Psychology of the Other One*. Columbia, MO.: The Missouri Book Company Publishers.
MILL, J.S. 1843. *A System of Logic Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. Edición de J.M. Robson e introducción de R.F. McCrae. Londres: Routledge & Kegan Paul, 1981.
MILLÁN-PUELLES, A. 1967. *La estructura de la subjetividad*. Madrid: Rialp.
——. 1990. *Teoría del objeto puro*. Madrid: Rialp.
MILLER, G.A. 1952. "Finite Markov Processes in Psychology." *Psychometrika* 17: 149-167.
——. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information." *Psychological Review* 63: 81-97.
——. 1984. "Informavores," en F. Machlup y U. Mansfield, eds. 1984.
——. 1986. "Interview with George A. Miller," en B.J. Baars 1986.
MILLER, G.A. y FRICK, F.C. 1949. "Statistical Behavioristics and Sequences of Responses." *Psychological Review* 56: 311-324.
MILLER, G.A., GALANTER, E. y PRIBRAM, K. 1960. *Plans and the Structure of Behavior*. Nueva York, NY.: Holt.

MILLER, J.G. 1955. "Toward a General Theory for the Behavioral Sciences." *American Psychologist* 10 (9): 513-531.
MILLIKAN, R.G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA.: MIT Press.
——. 1989. "Biosemantics." *Journal of Philosophy* 86: 218-297. Reimpreso en S.P. Stich y T.A. Warfield, eds. 1994a.
——. 1990a. "Truth Rules, Hoverflies, and the Kripke-Wittgenstein Paradox." *Philosophical Review* 94: 323-353.
——. 1990b. "Speaking up for Darwin," en B. Loewer y G. Rey, eds. 1991.
——. 1993. *White Queen Psychology and Other Essays for Alice*. Cambridge, MA.: MIT Press.
——. "Biosemantics," en B. McLaughlin, A. Beckermann y S. Walter, eds.
MINSKY, M. 1959. "Some Methods of Artificial Intelligence and Heuristic Programming." *Proceedings of the Symposium on the Mechanisation of Thought Processes, National Physical Laboratory, Teddington, England, November 24-27, 1958*. Londres: Her Majesty's Stationery Office.
——. 1969. *Semantic Information Processing*. Cambridge, MA.: MIT Press.
——. 1986. *The Society of Mind*. Nueva York, NY.: Simon and Schuster.
MISCHEL, T. ed. 1971. *Cognitive Development and Epistemology*. Nueva York, NY.: Academic Press.
MOLIÈRE, J.B. de POQUELIN. 1673. *El enfermo imaginario*. Traducción de J. Millás-Raurell. Madrid: Alba, 1999.
MOORE, G.E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.
——. 1939. "Proof of an External World." *Proceedings of the British Academy* 25: 273-300. Reimpreso en G.E. Moore 1993: 147-70.
——. 1944. "A Reply to my Critics," en P.A. Schilpp, ed. 1944.
——. 1993. *G.E. Moore: Selected Writings*. Edición de T. Baldwin. Londres: Routledge.
MOORE, T.E., ed. 1973. *Cognitive Development and the Acquisition of Language*. Nueva York, NY.: Academic Press.
MORA, F., ed. 1995. *El problema cerebro-mente*. Madrid: Alianza.
MORA, J.A., ed. 1992a. *Balance y futuro del conductismo tras la muerte de B.F. Skinner*. Málaga: Edinford.
——. 1992b. "Las contradicciones internas en el conductismo skinneriano," en J.A. Mora, ed. 1992a.
MORGAN, C.L. 1896. "Animal Automatism and Consciousness." *The Monist* 7: 1-18.
MORRIS, E.K. y TODD, J.T. 1999. "Watsonian Behaviorism," en W. O'Donohue y R. Kitchener, eds. 1999.
MORTON, A. 1975. "Because He Thought He Had Insulted him." *Journal of Philosophy* 72: 5-15.
MOSTERÍN, J. 1984. *Conceptos y teorías en la ciencia*. Madrid: Alianza, 2008.
MOSTERÍN, J. y TORRETTI, R. 2002. *Diccionario de Lógica y Filosofía de la Ciencia*. Madrid: Alianza.
MOWRER, O.H. 1960. *Learning Theory and Behavior*. Nueva York, NY.: Wiley.
MOYA, C.J. 1994. "Las emociones y la naturalización de la intencionalidad." *Anales del Seminario de Metafísica* 28: 227-255.
MUNITZ, M.R. y UNGER, P.K., eds. 1974. *Semantics and Philosophy*. Nueva York, NY.: New York University Press.
MUÑOZ, J. y VELARDE, J., eds. 2000. *Compendio de Epistemología*. Madrid: Trotta.
MURCHISON, C., ed. 1928. *Psychologies of 1925. Powell Lectures in Psychological Theory*. Worcester, MA.: Clark University Press.
NAGEL, E. 1950. "Science and Semantic Realism." *Philosophy of Science* 17: 174-181.
——. 1961. *The Structure of Science*. Nueva York, NY.: Harcourt, Brace and World.
NAGEL, T. 1965. "Physicalism." *Philosophical Review* 74: 339-356.
——. 1974. "What Is It Like to Be a Bat?" *Philosophical Review* 83: 435-450. Reimpreso en N.J. Block, ed. 1980.
——. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

NAGORNYI, N.M. y MARCHENKOV, S.S. 2001. "Turing Machine," en Hazewinkel, M., ed. 1987-2002.
NEEDHAM , J. 1929. *The Skeptical Biologist (Ten Essays)*. Londres: Chatto & Windus.
NEISSER, U. 1967. *Cognitive Psychology*. Nueva York, NY.: Appleton-Century-Crofts.
——., ed. 1987. *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge: Cambridge University Press.
NELSON, R.J. 1975. "Behaviorism, Finite Automata, and Stimulus Response Theory." *Theory and Decision* 6: 249-267.
von NEUMANN, J. 1951. "The General and Logical Theory of Automata," en L.A. Jeffres, ed. 1951.
NEWELL, A. 1962. "Some Problems of Basic Organization in Problem-Solving Programs," en A. Yovitts, G.T. Jacobi y G.D. Goldstein, eds. 1962.
——. 1973. "Production Systems: Models of Control Structures," en W. Chase, ed. 1973.
——. 1982. "The Knowledge Level." *Artificial Intelligence* 18: 87-127.
——. 1986. "The Symbol Level and the Knowledge Level," en Z. Pylyshyn y W. Demopoulos, eds. 1986.
NEWELL, A., SHAW, J.C. y SIMON, H.A. 1957. "Empirical Explorations of the Logic Theory Machine: A Case Study in Heuristic." *Proceedings of the Western Joint Computer Conference (Institute of Radio Engineers)*: 218-230.
——. 1958. "Elements of a Theory of Human Problem Solving." *Psychological Review* 65: 151-166.
NEWELL, A. y SIMON, H.A. 1958. "Heuristic Problem Solving: The Next Advance in Operations Research." *Operations Research* 6: 1-10.
——. 1963. "Computers in Psychology," en R.D. Luce, R.R. Bush y E. Galanter, eds. 1963.
——. 1972. *Human Problem Solving*. Englewood Cliffs, NJ.: Prentice-Hall.
——. 1975. "Computer Science as Empirical Inquiry: Symbols and Search." *Communications of the Association for Computing Machinery* 19: 113-126. Reimpreso en J. Haugeland, ed. 1981.
NEWTON, I. 1687. *Principios matemáticos de la filosofía natural*. Traducción de A. Escohotado. Madrid: Tecnos, 1997.
——. 1704. *Óptica, o tratado de las reflexiones, refracciones, inflexiones y colores de la luz*. Traducción de C. Solís según la edición de 1727. Madrid: Alfaguara, 1977.
NICOLELIS, M.A.L. y RIBEIRO, S. 2009. "En busca del código neural." *Temas de Investigación y Ciencia* 57: 10-17.
NOCKS, L. 2007. *The Robot: The Life-Story of a Technology*. Westport, CT.: Greenwood.
NOË, A. 2004. *Action in Perception*. Cambridge, MA.: MIT Press.
NORMAN, D., ed. 1981. *Perspectives on Cognitive Science*. Norwood, NJ.: Ablex.
O'CALLAGHAN, J.P. 1997. "The Problem of Language and Mental Representation in Aristotle and St. Thomas." *The Review of Metaphysics* 50: 499-541.
O'DONOHUE, W. y KITCHENER, R., eds. 1999. *Handbook of Behaviorism*. San Diego: Academic Press.
OETTINGER, A.G. 1952. "Programming a Digital Computer to Learn." *Philosophical Magazine* 43: 1243-1263.
OGDEN, C.K. y RICHARDS, I.A. 1923. *El significado del significado. Una investigación acerca de la infuencia del lenguaje sobre el pensamiento y de la ciencia simbólica*. Traducción de E. Prieto. Barcelona: Paidós, 1984.
OLDS, J. y MILNER, P. 1954. "Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain." *Journal of Comparative and Physiological Psychology* 47: 419–427. Reimpreso en T.K. Landauer, ed. 1967.
ORTONY, A., ed. 1979. *Metaphor and Thought*. Cambridge: Cambridge University Press.
OWEN, G.E.L. 1960. "Eleatic questions." *Classical Quarterly* 10: 84-102.
PALERMO, D.S. 1971. "Is a Scientific Revolution Taking Place in Psychology?" *Science Studies* 1: 135-155.
PALMER, S.E. 1999. "Color, Consciousness, and the Isomorphism Constraint." *Behavioral and Brain Sciences* 22(6): 923-943.

PAREDES, M.C. 2007. *Teorías de la intencionalidad*. Madrid: Síntesis.
PARRET, H. y BOUVERESSE, J., eds. 1981. *Meaning and Understanding*. Berlín: W. de Gruyter.
PASK, G. 1961. *An Approach to Cybernetics*. Londres: Hutchinson.
PAUSANIAS. *Descripción de Grecia*. Traducción de M.C. Herrero. Madrid: Gredos, 1994.
PERRY, R.B. 1917. "Purpose as Systematic Unity." *The Monist* 27: 352-375.
——. 1921. "A Behavioristic View of Purpose." *Journal of Philosophy* 18 (4): 85-105.
PIATELLI-PALMARINI, M., ed. 1980. *Language and Learning*. Cambridge, MA.: Harvard University Press.
PICCININI, G. 2004. "Functionalism, Computationalism, and Mental Contents." *Canadian Journal of Philosophy* 34 (3): 375-410.
——. 2007. "Computationalism, the Church-Turing Thesis, and the Church-Turing Fallacy." *Synthese* 154: 97-120.
PLACE, U.T. 1956. "Is Consciousness a Brain Process?" *British Journal of Philosophy* 47: 44-50.
——. 1999. "Ryle's Behaviorism," en W. O'Donohue y R. Kitchener, eds. 1999.
PLATÓN. *Diálogos*. Edición de E. Lledó *et al*. Madrid: Gredos, 1981-1998.
——. *Gorgias*. Traducción de J. Calonge. En Platón, *Diálogos*, II.
——. *Menón*. Traducción de F.J. Olivieri. En Platón, *Diálogos*, II.
——. *Crátilo*. Traducción de J.L. Calvo. En Platón, *Diálogos*, II.
——. *Fedro*. Traducción de E. Lledó. En Platón, *Diálogos*, III.
——. *República*. Traducción de C. Eggers Lan. En Platón, *Diálogos*, IV.
——. *Sofista*. Traducción de F. García. En Platón, *Diálogos*, V.
——. *Teeteto*. Traducción de F. García. En Platón, *Diálogos*, V.
——. *Timeo*. Traducción de F. Lisi. En Platón, *Diálogos*, VI.
PLAYFAIR, J. 1860. *Elements of Geometry; Containing the First Six Books of Euclid, with Two Books on the Geometry of Solids. To which are Added, Elements of Plane and Spherical Trigonometry*. Philadelphia, PA.: J.B. Lippincott & Co.
PEACOCKE, C. 1979. *Holistic Explanation*. Oxford: Oxford University Press / Clarendon Press.
PEIRCE, C.S. 1932. *Principles of Philosophy and Elements of Logic. Collected Papers of Charles Sanders Peirce, vols. 1-2.* Edición de C. Hartshorne y P. Weiss. Cambridge, MA.: Harvard University Press / Belknap Press.
PERLER, D. 1996. *Repräsentation bei Descartes*. Francfurt: Klostermann.
PILLSBURY, W.B. 1911. *Essentials of Psychology*. Nueva York, NY.: Macmillan.
——. 1929. *The History of Psychology*. Nueva York, NY.: Norton.
PIMENTEL, J. 2010. *El Rinoceronte y el Megaterio. Un ensayo de morfología histórica*. Madrid: Abada.
POLITZER, G. 1969. *Crítica de los fundamentos de la psicología.* Barcelona: Martínez Roca.
POPPER, K.R. 1972. *Conocimiento objetivo: un enfoque evolucionista*. Traducción de C. Solís. Madrid: Tecnos, 2005.
POPPER, K.R. y ECCLES, J.C. 1984. *El yo y su cerebro*. Traducción de C. Solís. Barcelona: Labor, 1985.
POSTMAN, L. 1963. "Does Interference Theory Predict Too Much Forgetting?" *Journal of Verbal Learning and Verbal Behavior* 2: 40-48.
PRADES, J.L. 1993. "Epistemología del contenido y del significado," en V. Sanfélix, ed. 1993.
PRESLEY, C.F., ed. 1967. *The Identity Theory of Mind*. Queensland: University of Queensland Press.
PRICE, H.H. 1953. *Thinking and Experience*. Londres: Hutchinson.
PRIOR, A.N. 1949. *Logic and the Basis of Ethics*. Oxford: Oxford University Press.
PSILLOS, S. 1999. *Scientific Realism: How Science Tracks Truth*. Londres: Routledge.
——. 2006. "Ramsey's Ramsey Sentences," en M. C. Galavotti, ed., 2006.
PUJADAS, L.M. 2002. *La ascensión y la caída de la teoría funcionalista de la mente*. Palma de Mallorca: Universitat de les Illes Balears.
PUTNAM, H. 1960. "Minds and Machines," en S. Hook, ed. 1960. Reimpreso en H. Putnam 1975c. Traducción de P. Navarro: "Mentes y máquinas," en A.M. Turing, H. Putnam y D. Davidson 1985.

——. 1962. "The Analytic and the Synthetic," en H. Feigl y G. Maxwell, eds. 1962. Reimpreso en H Putnam 1975d.
——. 1963a. "Brains and Behavior," en R.J. Butler, ed. 1963. Reimpreso en H. Putnam 1975c.
——. 1963b. "Degree of Confirmation and Inductive Logic," en P.A. Schilpp, ed. 1963. Reimpreso en H. Putnam 1975c.
——. 1967a. "The Nature of Mental States." Publicado como "Psychological Predicates" en W.H Capitan y D.D Merrill, eds. 1967. Reimpreso en H. Putnam 1975d. Traducción de M.M. Valdés: "La naturaleza de los estados mentales." *Cuadernos de Crítica* 15.
——. 1967b. "The Mental Life of Some Machines," en H.N. Castañeda, ed. 1967. Reimpreso en H. Putnam 1975d. Traducción de M. Gorostiza: "La vida mental de algunas máquinas." *Cuadernos de Crítica* 17.
——. 1975a. "The Meaning of 'Meaning'." En K. Gunderson, ed. 1975. Reimpreso en H. Putnam 1975d. Traducción de J. Flematti: "El significado de 'significado'." *Cuadernos de Crítica* 28. También en *Teorema* 14: 345-406 y L.M. Valdés Villanueva, ed. 2005.
——. 1975b. "Philosophy and Our Mental Life," en H. Putnam 1975d.
——. 1975c. *Mathematics, Matter, and Method. Philosophical Papers, vol. 1*. Cambridge: Cambridge University Press.
——. 1975d. *Mind, Language, and Reality. Philosophical Papers, vol. 2*. Cambridge: Cambridge University Press.
——. 1980. "What is Innate, and Why," en M. Piatelli-Palmerini, ed. 1980.
——. 1981. *Razón, verdad e historia*. Traducción de J.M. Esteban Cloquell. Madrid: Tecnos, 1988.
——. 1983a. *Realism and Reason. Philosophical Papers, vol. 3*. Cambridge: Cambridge University Press.
——. 1983b. "Computational Psychology and Interpretation Theory," ponencia presentada en Conference on the Foundations of Cognitive Science (University of Western Ontario, 1981), en H. Putnam 1983a.
——. 1985. "Reflexive Reflections." *Erkenntnis* 22: 143-153. Reimpreso en S. Silvers, ed. 1989.
——. 1986. "Meaning Holism," en L. Hahn y P. Schilpp, eds. 1986
——. 1997. "Functionalism. Cognitive Science or Science Fiction?" en D. Martel Johnson y C.E. Erneling, eds. 1997.
——. 1988. *Representation and Reality*. Cambridge, MA.: MIT Press.
PYLYSHYN, Z. 1979. "Complexity and the Study of Artificial and Human Intelligence," en M. Ringle, ed. 1979. Reimpreso en J. Haugeland, ed. 1981.
——. 1980. "Computation and Cognition: Issues in the Foundation of Cognitive Science." *Behavioral and Brain Sciences* 3: 111-132.
——. 1984. *Computación y conocimiento: hacia una fundamentación de la ciencia cognitiva*. Traducción de R. Fernández. Madrid: Debate, 1988.
PYLYSHYN, Z y DEMOPOULOS, W., eds. 1986. *Meaning and Cognitive Structure*, Norwood, NJ.: Ablex.
QUINE, W.v.O. 1953. "Dos dogmas del empirismo," traducción de M. Sacristán en L.M. Valdés Villanueva, ed. 2005.
——. 1956. "Quantifiers and Propositional Attitudes." *Journal of Philosophy* 53: 177-187. Reimpreso en W.v.O. Quine 1966.
——. 1960. *Palabra y objeto*. Traducción de M. Sacristán. Barcelona: Herder, 2001.
——. 1966. *The Ways of Paradox and Other Essays*. Nueva York, NY.: Random House. Segunda edición, revisada y ampliada: 1976. Cambridge, MA.: Harvard University Press.
——. 1969. *La relatividad ontológica y otros ensayos*. Traducción de M. Garrido y J.L. Blasco. Madrid: Tecnos, 1974.
——. 1974. *The Roots of Reference*. Chicago / La Salle, IL.: Open Court.
——. 1975. "Mind and Verbal Dispositions," en S. Guttenplan, ed. 1975.
——. 1979. "Intensions Revisited," en P.A. French, T.E. Uehling, y H.K. Wettstein, eds. 1979.
——. 1985. "States of Mind." *Journal of Philosophy* 82: 5-8.
——. 1990. *Pursuit of Truth*. Cambridge, MA.: Harvard University Press.
QUINTANA, J. 1985. *La psicología de la conducta: análisis crítico*. Madrid: Alhambra.
RABOSSI, E. 1995. "La tesis de la identidad mente-cuerpo," en F. Broncano, ed. 1995.

RABOSSI, E., ed. 1996. *Filosofía y Ciencia Cognitiva*. Barcelona: Paidós.
RAMÓN Y CAJAL, S. 1897/1941. *Los tónicos de la voluntad. Reglas y consejos sobre investigación científica*. Edición de L. López-Ocón. Madrid: Gadir, 2006.
RAMSEY, F. 1931. *The Foundations of Mathematics and Other Essays*. Edición de R. B. Braithwaite. Londres: Routledge & Kegan Paul.
RAPAPORT, D., ed. 1951. *Organization and Pathology of Thought*. Nueva York, NY.: Columbia University Press.
RASHEVSKY, N. 1931. "Possible Brain Mechanisms and their Physical Models." *Journal of General Psychology* 5: 368-406.
REED, E. 1997. "The Cognitive Revolution from an Ecological Point of View," en D. Martel Johnson y C.E. Erneling, 1997, eds.
REID, T. 1785. *Essays on the Intellectual Powers of Man*. Hildesheim: G. Olms 1967. También en *The Edinburgh Edition of Thomas Reid*, III. University Park: Pennsylvania State University Press, 2002.
REINES, F. y COWAN Jr., C.L. 1956. "The Neutrino." *Nature* 178: 446.
RESCHER, N. 2001. *Paradoxes: Their Roots, Range, and Resolution*. Chicago / La Salle, IL.: Open Court.
REYES, R., ed. 2009. *Diccionario Crítico de Ciencias Sociales*. Madrid / México, D.F.: Plaza y Valdés.
RICHARDS, I.A. 1936. *The Philosophy of Rhetorics*. Oxford: Oxford University Press.
RICHARDSON, R. 1979. "Functionalism and Reductionism." *Philosophy of Science* 46: 533-558.
RIEMANN, B. 1867. "Über die Hypothesen, welche der Geometrie zu Grunde liegen." *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen* 13: 143-152. Editado por R. Dedekind.
RINGEN, J. 1990. "Radical Behaviorism: B.F. Skinner's Philosophy of Science," en W. O'Donohue y R. Kitchener, eds. 1990.
RINGLE, M., ed. 1979. *Philosophical Perspectives in Artificial Intelligence*. Atlantic Highlands, NJ.: Humanities Press.
RIPS, L. 1989. "Similarity, typicality, and categorization," en S. Voisniadou y A. Ortony, eds. 1989.
RIVIÈRE, Á. 1977. "El análisis experimental de la conducta y el conductismo radical como filosofía." *Investigación y Ciencia* 7: 114-118. Reimpreso en Á. Rivière 2002.
——. 1991a. *Objetos con mente*. Madrid: Alianza.
——. 1991b. "Orígenes históricos de la psicología cognitiva: paradigma simbólico y procesamiento de la información." *Anuario de Psicología* 51: 129-155.
——. 1993. "Las multitudes de la mente." *Anuario de Psicología* 56: 112-144. Reimpreso en Á. Rivière 2002.
——. 1995. "Mentes, cerebro y cómputos: ¿problemas o misterios?," en F. Mora, ed. 1995. Reimpreso en Á. Rivière 2002.
——. 2002. *Obras escogidas, I. Diálogos sobre Psicología: de los cómputos mentales al significado de la conciencia*. Madrid: Editorial Médica Panamericana.
ROBLES, J.A. y SILVA, C. 1956. Prólogo a John Locke, *Ensayo sobre el entendimiento humano*.
ROCHESTER, N., HOLLAND, J.H., HAIBT, L.H. y DUDA, W.L. 1956. "Test on a Cell Assembly Theory of the Action of the Brain, Using a Large Digital Computer." *IRE Transactions of Information Theory* IT-2 (3): 80-93.
RODRÍGUEZ, M. 2001a. "El conductismo lógico," en P. Chacón *et al.* 2001.
——. 2001b. "Intencionalidad y contenido mental," en P. Chacón *et al.* 2001.
——. 2005a. *La mente en sus máscaras. Ensayos de filosofía de la psicología*. Madrid: Biblioteca Nueva.
——. 2005b. "El interés de la teoría kantiana de la mente para la ciencia cognitiva: una contribución introductoria," en M. Rodríguez, ed. 2005a.
de ROJAS, C. 1598. *Teoría y práctica de fortificación, conforme las medidas y defensas destos tiempos*. Madrid: Luis Sánchez.
RORTY, A.O., ed. 1976. *The Identities of Persons*. Berkeley, CA.: University of California Press.

RORTY, R. 1965. "Mind-Body Identity, Privacy, and Categories." *Review of Metaphysics* 19: 24-54.
——. 1979. *La filosofía y el espejo de la naturaleza*. Traducción de J. Fernández Zulaica: Madrid: Cátedra, 1989.
ROSCH, E. 1973. "On the internal structure of perceptual and semantic categories," en T.E. Moore, ed. 1973.
——. 1975. "Cognitive representation of semantic categories." *Journal of Experimental Psychology: General* 104: 192-233.
——. 1978. "Principles of categorization," en E. Rosch y B. Lloyd, eds. 1978.
ROSCH, E. y LLOYD, B., eds. 1978. *Cognition and Categorization*. Hillsdale, NJ.: Erlbaum.
ROSENTHAL, D.M. 1986. "Intentionality," en P.A. French, T.E. Uehling y H.K. Wettstein, eds. 1986. Reimpreso en S. Silvers, ed. 1989.
——, ed. 1991. *The Nature of Mind*. Oxford: Oxford University Press.
——. 2005. *Consciousness and Mind*. Oxford: Oxford University Press / Clarendon Press.
ROSS, T. 1933. "Machines that Think." *Scientific American* 148: 206-208.
——. 1935. "Machines that Think. A Further Statement." *Psychological Review* 42: 387-393.
——. 1938. "The Synthesis of Intelligence. Its Implications." *Psychological Review* 45: 185-189.
RUSSELL, A. 1953. *Charles M. Russell. Cowboy Artist*. Nueva York: Twayne.
RUSSELL, B. 1903. *Principles of Mathematics*. Cambridge: Cambridge University Press.
——. 1908. "Mathematical Logic as Based on the Theory of Types," *American Journal of Mathematics* 30: 222–262. Reimpreso en B. Russell. *Logic and Knowledge*. Londres: Allen and Unwin, 1956.
——. 1911. "Knowledge by Acquaintance and Knowledge by Description." *Proceedings of the Aristotelian Society* 11: 108-128.
——. 1927. *Fundamentos de Filosofía*. Traducción de R. Crespo y Crespo. Barcelona: Apolo, 1936 / Plaza & Janés, 1985.
——. 1940. *Investigación sobre el significado y la verdad*. Traducción de J. Rovira Armengol. Buenos Aires: Losada, 2003.
——. 1945. *Historia de la Filosofía occidental* (2 vols.). Traducción de J. Gómez de la Serna y A. Dorta. Buenos Aires: Espasa-Calpe, 1947.
RUSSELL, S. y NORVIG, R., eds. 2003. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ.: Prentice Hall.
RYLE, G. 1949. *El concepto de lo mental*. Traducción de E. Rabossi. Barcelona: Paidós, 2005.
SAHLIN, N.E. 1990. *The Philosophy of F.P. Ramsey*. Cambridge: Cambridge University Press.
SAMET, J. y FLANAGAN, O.J. 1989. "Innate Representations," en S. Silvers, ed. 1989.
SANFÉLIX, V., ed. 1993. *Acerca de Wittgenstein*. Valencia: Pretextos.
SAVAGE, C.W., ed. 1978. *Perception and Cognition: Issues in the Foundations of Psychology (Minnesota Studies in the Philosophy of Science IX)*. Minneapolis, MN.: University of Minnesota Press.
SAYRE, K. 1986. "Intentionality and Information Processing: An Alternative Model for Cognitive Science." *Behavioral and Brain Sciences* 9 (1): 121-138.
SCARBOROUGH, D., OSHERSON, D.N. y STERNBERG, S., eds. 1997. *An Invitation to Cognitive Science, vol. 4*. Cambridge, MA.: MIT Press.
SCHANK, R. C. 1972. "Conceptual Dependence: A Theory of Natural Language Understanding." *Cognitive Psychology* 3: 552-631.
SCHEERER, E. 1987. "The Unknown Fechner." *Psychological Research* 49: 197-202.
SCHIFFER, S. 1972. *Meaning*. Oxford: Oxford University Press / Clarendon Press.
——. 1978. "The Basis of Reference." *Erkenntnis* 13: 171-206.
——. 1981. "Truth and the Theory of Content," en H. Parret y J. Bouveresse, eds. 1981.
——. 1982. "Commentary on R. Matthews' *Knowledge of Language in a Theory of Language Processing*". Ponencia presentada en la Conferencia *Constraints on Modeling Real-Time Processes*. Cape Camargue, Francia, junio de 1982.
——. 1986. "Functionalism and Belief," en M. Brand y R.M. Harnish, eds. 1986.
——. 1987. *Remnants of Meaning*. Cambridge, MA.: MIT Press.
——. 1991. "Ceteris Paribus Laws." *Mind* 100 (1/397): 1-17.

SCHILPP, P.A., ed. 1944. *The Philosophy of G.E. Moore*. Chicago / La Salle, IL.: Open Court.
——. 1963. *The Philosophy of Rudolf Carnap*. Chicago / La Salle, IL.: Open Court.
SCHLICK, M. 1918. *Allgemeine Erkenntnislehre*. Berlín: Springer. Segunda edición revisada, 1925.
SCRIVEN, M. 1956. "A Study of Radical Behaviorism," en H. Feigl y M. Scriven, eds. 1956.
SILVERS, S., ed. 1989. *Rerepresentation. Readings in the Philosophy of Mental Representation*. Dordrecht: Kluwer.
SIMON, H.A. y NEWELL, A. 1956. "Models: their Uses and Limitations," en L.D. White, ed. 1956.
SEARLE, J.R. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge: Cambridge University Press.
——. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3: 417-424.
——. 1982. "The Myth of the Computer (Review of *The Mind's I: Fantasies and Reflections on Self and Soul*, Composed and Arranged by D. Hofstadter and D.C. Dennett)." *The New York Review of Books* 29 (7): 3-6.
——. 1983. *Intentionality. An Essay in the Philosophy of Mind*. Londres: Cambridge University Press.
——. 1989. "Conciousness, Unconsciousness, and Intentionality." *Philosophical Topics* 17: 193-209.
——. 1990a. "Consciousness, Explanatory Inversion, and Cognitive Science." *Behavioral and. Brain Sciences* 13(3): 585-596
——. 1990b. *The Mistery of Consciousness*. Nueva York, NY.: The New York Review of Books.
——. 1990c. "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64: 21-37.
——. 1990d. "Is the Brain's Mind a Computer Program?" *Scientific American* 262 (1): 20-25
——. 1992. *El redescubrimiento de la mente*. Traducción de L.M. Valdés-Villanueva. Barcelona: Crítica, 1996.
SEDIVY, S. 1990. *The Determinate Character of Perceptual Experience*. Tesis doctoral inédita, *apud* McDowell 1994a: 111, nota. Pittsburgh: Universidad de Pittsburgh.
SELLARS, W. 1956. *Empiricism and the Philosophy of Mind*. Minneapolis, MN.: University of Minnesota Press. Reeditado con introducción de R. Rorty y guía para el estudio de R. Brandom. Cambridge, MA.: Harvard University Press, 1997.
——. 1963. *Ciencia, percepción y realidad*. Traducción de V. Sánchez de Zabala. Madrid: Tecnos, 1971.
——. 1969. "Language as Thought and as Communication." *Philosophy and Phenomenological Research* 29: 506-527.
SÉNECA. *Cartas morales a Lucilio*. Edición y traducción de E. Sierra. Madrid: Planeta, 1985.
——. *Sobre la ira*. Traducción de F. Navarro. La Laguna: Artemisa, 2007.
SHAFFER, J. 1961. "Could Mental States Be Brain Processes?" *Journal of Philosophy* 58: 813-822.
SHANNON, C.E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27: 379-423.
SHAPIRO, L. 2000. "Multiple Realizations." *Journal of Philosophy* 97: 635-654.
——. 2004. *The Mind Incarnate*. Cambridge, MA.: MIT Press.
SHEPARD, R. 1987. "Toward a Universal Law of Generalization for Psychological Science." *Science* 237: 1317-1323.
SHOEMAKER, S. 1975. "Functionalism and Qualia." *Philosophical Studies* 27: 271-315.
——. 1982. "The Inverted Spectrum." *Journal of Philosophy* 79: 357-381.
——. 1992. "The First-Person Perspective," en N.J. Block, O.J. Flanagan y G. Güzeldere, eds. 1997.
SHOTTER, J. 1997. "Cognition as a Social Practice. From Computer Power to Word Power," en D. Martel Johnson y C.E. Erneling, eds. 1997.
SIGUÁN, M. 1993. "Objetos con mente como sujetos de la psicología." *Anuario de Psicología* 56: 76-84.

SIMMEL, G. 1892. *Problemas de Filosofía de la Historia*. Traducción de Elsa Tabernig. Buenos Aires: Nova, 1950.
——. 1918. *Vom Wesen des historischen Verstehens*. Berlín: E.S. Mittler & Sohn.
SIMON, H.A. 1969. *The Sciences of the Artificial*. Cambridge, MA.: MIT Press.
SIMON, H.A. y SIKLÓSSY, L., eds. 1972. *Representation and Meaning*. Englewood Cliffs, NJ.: Prentice-Hall.
SIROTIN, Y.B. y DAS, A. 2009. "Anticipatory Haemodynamic Signals in Sensory Cortex Not Predicted by Local Neuronal Activity." *Nature* 457: 475-479.
SKINNER, B.F. 1931. "The Concept of the Reflex in the Description of Behavior." *Journal of General Psychology* 5: 427-458. Traducción en B.F. Skinner 1975.
——. 1932. "Drive and Reflex Strenght." *Journal of General Psychology* 6: 22-37 / 38-48.
——. 1935. "The Generic Nature of the Concepts of Stimulus and Response." *Journal of General Psychology* 12: 40-65.
——. 1938. *The Behavior of Organisms: An Experimental Analysis*. Nueva York, NY.: Appleton-Century-Crofts.
——. 1945. "The Operational Analysis of Psychological Terms." *Psychological Review* 52: 270-272/291-294. Traducción castellana en B.F. Skinner, 1975.
——. 1953. *Ciencia y conducta humana.* Traducción de M.J. Gallofré. Barcelona: Fontanella, 1971.
——. 1957. *Verbal Behavior*. Nueva York, NY.: Appleton-Century-Crofts.
——. 1971. *Más allá de la libertad y la dignidad*. Traducción de J.J. Coy. Barcelona: Fontanella, 1972.
——. 1974. *Sobre el conductismo*. Traducción de F. Barrera, revisión y prólogo de R. Ardila. Barcelona: Fontanella, 1975.
——. 1975. *Registro acumulativo: selección de la obra de B.F. Skinner realizada por el propio autor*. Traducción de R. Berdagué Costa. Barcelona: Fontanella.
——. 1977. "Why I am not a Cognitive Psychologist." *Behaviorism* 5 (2): 1-11.
——. 1979. *The Shaping of a Behaviorist*. Nueva York, NY.: Alfred A. Knopf.
——. 1985. "Cognitive Science and Behaviourism." *British Journal of Psychology* 76: 291-301.
——. 1987. "Whatever Happened to Psychology as the Science of Behavior?" *American Psychologist* 8: 780-786.
——. 1989. "The Origins of Cognitive Thought." *American Psychologist* 10: 13-18.
——. 1990. "Can Psychology Be a Science of the Mind?" *American Psychologist* 11:1206-1210.
SMART, J.C.C. 1959. "Sensations and Brain Processes." *Philosophical Review* 68: 141-156.
——. 1967. "Comments on the Papers," en C.F. Presley, ed. 1967.
——. 1972. "Further Thoughts on the Identity Theory." *The Monist* 56: 149-162.
SMITH, B.C. 1996. *On the Origin of Objects*. Cambridge, MA.: MIT Press.
——. 2002a. "Cummins –or Something Isomorphic to Him." En H. Clapin, ed. 2002a.
——. 2002b. "Reply to Dennett." En H. Clapin, ed. 2002a.
SMITH, E.E., y OSHERSON, D.N., eds. 1990. *An Invitation to Cognitive Science, vol. 3*: *Thinking*. Cambridge, MA.: MIT Press. Segunda edición revisada: 1995.
SMITH, L.D. 1986. *Behaviorism and Logical Positivism. A Reassessment of the Alliance*. Stanford: Stanford University Press.
SMOLENSKY, P. 1988, "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences* 11: 1-23.
SMULLYAN, R. 1978. *¿Cómo se llama este libro? El enigma de Drácula y otros pasatiempos lógicos*. Traducción de L.M. Valdés-Villanueva. Madrid: Cátedra, 1989.
SOBER, E. 1999. "The Multiple Realizability Argument Against Reductionism." *Philosophy of Science* 66: 542-564.
SOKOLOWSKI, R. 1987. "Exercising Concepts." *The Review of Metaphysics* 40: 451-463.
SORENSEN, R. 1988. *Blindspots*. Oxford: Oxford University Press / Clarendon Press.
——. 2003. *Breve historia de la paradoja. La filosofía y los laberintos de la mente*. Traducción de A. E. Álvarez y R. Orsi. Barcelona: Tusquets, 2007.
SOSA, E. 1970. "Propositional Attitudes *de Dicto* and *de Re*." *Journal of Philosophy* 67: 883-896.
——, ed. 1975. *Causation and Conditionals*. Oxford: Oxford University Press.
——. 1993. "Davidson's Thinking Causes," en J. Heil y A. Mele, eds. 1993.

de SOUSA, R. 1987. *The Rationality of Emotion*. Cambridge, MA.: MIT Press.
SPERBER, D. ed. 2000. *Metarepresentations: A Multidisciplinary Perspective*. Oxford: Oxford University Press / Clarendon Press.
SQUIRES, R. 1968. "Are Dispositions Causes?" *Analysis* 29 (2): 45-47.
STAHL, G.E. 1715. *Opusculum chymico-physico-medicum, seu schediasmatum a pluribus annis variis occasionibus in publicum emissorum nunc... in unum volumen jam collectorum...* Halae Magdeburgicae [Halle]: typis et impensis Orphanotrophei.
STALNAKER, R. 1976. "Propositions," en A. McKay y D.D. Merrill, eds. 1976.
——. 1984. *Inquiry*. Cambridge, MA.: MIT Press.
STAMPE, D. 1975. "Show and Tell," en B. Freed *et al.*, eds. 1975.
——. 1977. "Toward a Causal Theory of Linguistic Representation," en P.A. French, T.E. Uehling y H.K. Wettstein, eds. 1977.
STEPHENS, J.M. 1929. "A Mechanical Explanation of the Law of Effect." *American Journal of Psychology* 41: 422-431.
STEVENSON, J.T. 1960. "'Sensations and Brain Processes': A Reply to J.J.C. Smart," *Philosophical Review* 69: 505-510. Reimpreso en C.V. Borst, ed. 1970.
STICH, S.P. 1978. "Autonomous Belief and the Belief-Desire Thesis." *The Monist* 61: 573-591.
——. 1983. *From Folk Psychology to Cognitive Science: the Case Against Belief*. Cambridge, MA.: MIT Press.
——. 1992. "What is a Theory of Mental Representation?" *Mind* 101: 243-261. Reimpreso en S.P. Stich y T.A. Warfield, eds. 1994a.
——. 1996. *Deconstructing the Mind*. Oxford: Oxford University Press.
STICH, S.P. y WARFIELD, T.A., eds. 1994a. *Mental Representation. A Reader*. Oxford: Blackwell.
——. 1994b. "Introduction," en S.P. Stich y T.A. Warfield, eds. 1994a.
STRAWSON, G. 1994. *Mental Reality*. Cambridge, MA.: MIT Press.
STRAWSON, P.F. 1950. "On Referring." *Mind* 59: 320-344. Traducción de L.M. Valdés Villanueva: "Sobre el referir", en L.M. Valdés Villanueva, ed. 1999.
——. 1952. *Introduction to Logical Theory*. Nueva York, NY.: Wiley.
——. 1979. "Perception and its Objects," en G.F. McDonald, ed. 1979.
STROUD, B. 2000. *The Quest for Reality: Subjectivism and the Metaphysics of Colour*. Oxford: Oxford University Press / Clarendon Press.
SUPPES, P., HENKIN, L., MOISIL, G.C. y JOJA, A., eds. 1973. *Proceedings of the Fourth International Congress for Logic, Methodology, and Philosophy of Science (Bucharest, 1971)*. Amsterdam: North Holland / Elsevier.
SUR, M. 2004. "Rewiring Cortex: Cross-Modal Plasticity and its Implications for Cortical Development and Function", en G.A. Calvert, C. Spence y B.E. Stein, eds. 2004.
TAYLOR, C. 1964. *The Explanation of Behaviour*. Nueva York, NY.: Humanities Press.
——. 1989. *Fuentes del yo. La construcción de la identidad moderna*. Traducción de A. Lizón. Barcelona: Paidós, 1996.
TENNANT, N. 2007. "Mind, Mathematics, and the *Ignorabimusstreit*." *British Journal for the History of Philosophy* 15 (4): 745-773.
THAGARD, P. 1992. *Conceptual Revolutions*. Princeton: Princeton University Press.
——. 2005. *La mente. Introducción a las ciencias cognitivas*. Traducción de S. Jawerbaum y J. Barba. Buenos Aires: Katz, 2008.
——, ed. 2007. *Handbook of the Philosophy of Science. Philosophy of Psychology and Cognitive Science*. Amsterdam: North Holland / Elsevier.
THORNDIKE, E.L. 1905. *The Elements of Psychology*. NuevaYork: Seiler. Segunda edición: 1912.
——. 1911. *Animal Intelligence: Experimental Studies.* Nueva York, NY.: Hafner.
——. 1927. "The Law of Effect." *The American Journal of Psychology* 39: 212-222.
——. 1931. *Human Learning*. Nueva York, NY.: Macmillan.
TOLLIVER, J.T. 1989. "Beliefs out of Control," en S. Silvers, ed. 1989.
TOOLEY, M. 1990. "The Nature of Causation: A Singularist Account." *Canadian Journal of Philosophy* 16 (Supplement): 271-322.

TINBERGEN, N. 1951. *The Study of Instinct*. Oxford: Oxford University Press / Clarendon Press.
——. 1952. "The Curious Behavior of the Stickleback." *Scientific American*: 22-26.
TOLMAN, E.C. 1925. "Behaviorism and Purpose." *Journal of Philosophy* 22: 36-41.
——. 1927. "A Behaviorist Definition of Consciousness." *Psychological Review* 34 (1): 433-439.
——. 1932/1967. *Purposive Behavior in Animals and Men*. Nueva York, NY.: Appleton-Century-Crofts.
——. 1936. "Operational Behaviorism and Current Trends in Psychology." Reimpreso como "The Intervening Variable" en M.H. Marx, ed. 1951.
——. 1938. "The Determiners of Behavior at a Choice Point," *Psychological Review* 45(1):1-41.
——. 1939. "Prediction of Vicarious Trial and Error by Means of the Schematic Sowbug." *Psychological Review* 46(4): 318-336.
——. 1948. "Cognitive Maps in Rats and Men." *Psychological Review* 55(4): 189-208.
——. 1959. "Principles of Purposive Behavior," en S. Koch, ed. 1959b.
TORIBIO, J. 1991. "Causal Efficacy, Content and Levels of Explanation." *Logique et Analyse* 135/136: 297-318. URL = < http://www.era.lib.ed.ac.uk/bitstream/1842/1360/1/CausalEfficacy.pdf, 21 de junio de 2011.
——. 1995. "Eliminativismo y el futuro de la psicología popular," en F. Broncano, ed. 1995.
——. 1998. "Meaning and Other Non-Biological Categories." *Philosophical Papers* 27 (2): 129-150. URL = <http://www.era.lib.ed.ac.uk/bitstream/1842/1369/1/ MeanOtherB.pdf, 21 de junio de 2011.
——. 2000. "Internalismo, Externalismo y Ecología," en P. Chacón y M. Rodríguez, eds. 2000.
TURING, A.M. 1936. "On Computable Numbers, With an Application to the *Entscheidungsproblem*." *Proceedings of the London Mathematical Society* 42: 230-265.
——. 1937. "On Computable Numbers, With an Application to the *Entscheidungsproblem*. A Correction." *Proceedings of the London Mathematical Society* 43: 544-546.
——. 1946. "Proposal for the Development in the Mathematics Division of an Automatic Computing Engine (ACE)," en B.E. Carpenter y R.W. Doran, eds. 1986.
——. 1947. "Lecture to the London Mathematical Society on 20 February 1947," en B.E. Carpenter y R.W. Doran, eds. 1986.
——. 1948. "Intelligent Machinery." *National Physical Laboratory Report*, reimpreso en B. Meltzer y D. Michie, eds. 1969b. URL = < http://www.AlanTuring.net/intelligent_machinery, 1 de agosto de 2010.
——. 1950a. "Computing machinery and intelligence." *Mind* 59: 433-460. Traducción de P. Navarro: "¿Puede pensar una máquina?", en A.M. Turing, H. Putnam y D. Davidson 1985. Traducción de G. Feher de la Torre: "La maquinaria de computación y la inteligencia," en M. Boden, ed. 1990.
——. 1950b. "Programmers' Handbook for Manchester Electronic Computer." *University of Manchester Computing Laboratory*. URL = <http://www.AlanTuring.net/programmers_handbook, 1 de agosto de 2010.
TURING, A.M., PUTNAM, H. y DAVIDSON, D. 1985. *Mentes y máquinas*. Traducción de P. Navarro. Madrid: Tecnos.
TVERSKY, A. y KAHNEMAN, D. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185: 1124-1131.
TYNDALL, J. 1871. *Fragments of Science for Unscientific People*: *a Series of Detached Essays, Addresses and Reviews.* Nueva York, NY.: Appleton.
URIAGEREKA, J. 2008. *Syntactic Anchors: on Semantic Structuring*. Cambridge: Cambridge University Press.
VALDÉS VILLANUEVA, L.M., ed. 2005. *La búsqueda del significado*. Madrid: Tecnos. Cuarta edición ampliada.
de VAUCANSON, J. 1738. "Lettre à M. l'Abbé Desfontaines," reimpreso parcialmente como "Relación sobre el mecanismo de un autómata. Carta de Jacques de Vaucanson al abad de Fontaine (1738)," traducción de J.C. Vales, en S. Bueno y M. Peirano, eds. 2009.

de VEGA, M. 1981. "Una exploración de los metapostulados de la psicología contemporánea: el logicismo." *Análisis y Modificación de Conducta* 7 (16): 345-375.
VELARDE-MAYOL, V. 2007. "El objeto puro en Meinong." *Diánoia* LII (58): 27-48.
VENDLER, Z. 1972. *Res Cogitans*. Ithaca, NY.: Cornell University Press.
VERPLANCK, W.S. 1954. "Burrhus F. Skinner," en W.K. Estes *et al*., eds. 1954.
VOISNIADOU, S. y ORTONY, A. *Similarity, Analogy, and Thought*. Cambridge: Cambridge University Press.
WAGNER, A.R. 1969. "Stimulus Validity and Stimulus Selection in Associative Learning," en N.J. McKintosh y W.K. Honig, eds. 1969.
WAGNER, S. y WARNER, R., eds. 1993. *Naturalism: A Critical Appraisal*. Notre Dame, IN.: University of Notre Dame Press.
WALLACE, R.A. 1952. "The Maze-Solving Computer." *Proceedings of the Association for Computing Machinery 1952 National Meeting*: 119-125.
WALTER, S. y HECKMANN, H., eds. 2003. *Physicalism and Mental Causation*. Charlottesville, VA.: Imprint Academic.
WARREN, N. 1971. "Is a Scientific Revolution Taking Place in Psychology? Doubts and Reservations." *Science Studies* 1: 407-413.
WARRINGTON, E.K. y TAYLOR, A.M. 1973. "The Contribution of the Right Parietal Lobe to Object Recognition." *Cortex* 9: 152-164.
——. 1978. "Two Categorical Stages of Object Recognition." *Perception* 7: 695-705.
WASON, P.C. 1966. "Reasoning," en B.M. Foss, ed. 1966.
——. 1968. "Reasoning about a Rule." *Quarterly Journal of Experimental Psychology* 20: 273-281.
WATSON, J.B. 1907. "Kinaesthetic and Organic Sensations: Their Role in the Reactions of the White Rat to the Maze." *Psychological Review Monograph Supplement* 8 (33): 1-100.
——. 1913. "Psychology as the Behaviorist Views It." *Psychological Review* 20: 158-177.
——. 1914. *Behavior: An Introduction to Comparative Psychology*. Nueva York, NY.: Holt.
——. 1924. *Behaviorism*. Nueva York, NY.: People's Institute. Segunda edición revisada: 1930. Chicago, IL.: University of Chicago Press. Nueva edición, con introducción de G.A. Kimble: 1998. New Brunswick, NJ.: Transaction Publishers.
WEBER, M. 1922. *Wirtschaft und Gesellschaft*. *Grundriss der Sozialökonomik*, III. Traducción parcial al inglés de A.M. Henderson y T. Parsons: *The Theory of Social and Economic Organization*. Oxford / Nueva York, NY.: Oxford University Press, 1947. Traducción al castellano de J. Ferrater Mora: *Economía y Sociedad: Esbozo de Sociología Comprensiva*. México, D.F.: Fondo de Cultura Económica, 1964.
——. 1984. *La acción social: ensayos metodológicos*. Traducción de M. Faber-Kaiser y S. Giner. Barcelona: Península.
WECHSLER, L. 1995. *Mr. Wilson's Cabinet of Wonder. Pronged Ants, Horned Humans, Mice on Toast and Other Marvels of Jurassic Technology*. Nueva York, NY.: Pantheon.
WEIMER, W.B. y PALERMO, D.S. 1973. "Paradigms and Normal Science in Psychology." *Science Studies* 3: 211-244.
——, eds. 1974. *Cognition and the Symbolic Processes*. Hillsdale, NJ.: Erlbaum.
WEISS, A.P. 1924. "Behaviorism and behavior." *Psychological Review* 29: 329-344.
——. 1925. *A Theoretical Basis of Human Behavior*. Segunda edición revisada: 1929. Columbus, OH.: R.G. Adams & Co. Reimpresión con estudio introductorio de R.H. Wozniak: 1994. Londres: Routledge / Thoemess Press y Tokio: Kinokuniya.
WEIZENBAUM, J. 1966. "ELIZA. A Computer Program for the Study of Natural Language Communication between Man and Machine." *Communications of the Association for Computing Machinery* 9(1): 36-45.
WEST, R. 2006. *Theory of Addiction*. Oxford: Blackwell.
WHITE, A. R. 1957. "On Claiming to Know." *Philosophical Review* 66 (2): 180-192.
WHITE, L.D., ed. 1956. *The State of the Social Sciences*. Chicago, IL.: Chicago University Press.
WHITE, S. 1986. "Curse of the Qualia", *Synthese* 68: 333-368.
——. 1999. "Why the Distinct Property Dualism Argument Won't Go Away," inédito. Ponencia presentada al New York University Language and Mind Colloquium, 4 de

abril de 1999, y al Workshop on Conceivability and Possibility de la Universidad de Friburgo.
WHYTE, J. 1990. "Success Semantics." *Analysis* 50: 149-157. URL = http:// www.nyu.edu/gsas/dept/philo/courses/consciousness/papers/WHYPDAW.html
WILKES, K. 1981. "Functionalism, Psychology, and Philosophy of Mind." *Philosophical Topics* 12: 147-167.
WILLIS, R. 1821. "An Attempt to Analyze the Automaton Chess Player of Mr. de Kempelen, with an Easy Method of Imitating the Movements of that Celebrated Figure, Illustrated by Original Drawings, to which Is Added a Copious Collection of the Knight's Moves over the Chess Board." Londres: Booth.
WILSON, R.A. y CRAVER, C.F. 2007. "Realization: Metaphysical and Scientific Perspectives," en P. Thagard, ed. 2007.
WILLIAMS, B. 1978. *Descartes. El proyecto de la investigación pura*. Traducción de J. Coll. Madrid: Cátedra, 1996.
WINCH, P. 1958. *The Idea of a Social Science and its Relation to Philosophy*. Londres: Routledge & Kegan Paul.
WINDELBAND, W. 1894. *Geschichte und Naturwissenschaft. Straßburger Rektoratsrede*, en *Präludien. Aufsätze und Reden zur Philosophie und ihrer Geschichte*. Tubinga: J.C.B. Mohr.
WINOGRAD, T. 1976. "Artificial Intelligence and Language Comprehension," en T. Winograd, ed. 1976.
WINOGRAD, T., ed. 1976. *Artificial Intelligence and Language Comprehension*. Washington, DC.: National Institute of Education.
——. 1981. "What Does it Mean to Understand Language?," en D. Norman, ed. 1981.
WIMSATT, W.C. 1972. "Teleology and the logical structure of function statements." *Studies in the History and Philosophy of Science* 3: 1-80.
WITMER, G. 2003. "Multiple Realizability and Psychological Laws: Evaluating Kim's Challenge," en S. Walter y H. Heckmann, eds. 2003.
WITTGENSTEIN, L. 1922. *Tractatus Logico-Philosophicus*. Traducción de J. Muñoz e I. Reguera. Madrid: Alianza, 2002.
——. 1946-1949. *Remarks on the Philosophy of Psychology*. Edición de G.E.M. Anscombre y G.H. von Wright. Chicago, IL.: University of Chicago Press, 1980.
——. 1953. *Investigaciones Filosóficas*. Traducción de A. García Suárez y U.C. Moulines:. Barcelona: Crítica, 1986.
——. 1958. *Los cuadernos azul y marrón*. Traducción de F. Gracia. Madrid: Tecnos, 2007.
——. 1969. *Sobre la certeza*. Edición de G. E. M. Anscombe y G. H. von Wright; traducción de Josep Lluís Prades y Vicent Raga. Barcelona: Gedisa, 1988.
WOOD, G. 2002. *Living Dolls. A Magical History of the Quest for Mechanical Life*. Londres: Faber & Faber / *Edison's Eve. A Magical History of the Quest for Mechanical Life*. Nueva York, NY.: Knopf.
WOODFIELD, A., ed. 1982. *Thought and Object. Essays on Intentionality*. Oxford: Oxford University Press / Clarendon Press.
WOODRIDGE, D. 1963. *The machinery of the brain*. Nueva York, NY.: McGraw Hill.
WOOZLEY, A.D. 1953. "Knowing and Not Knowing." *Proceedings of the Aristotelian Society* 53: 151-172.
WOZNIAK, R.H. 1993. "Max Meyer and *The Fundamental Laws of Human Behavior*," en M.F. Meyer (1911/1993).
——. 1994. "Albert Paul Weiss and *A Theoretical Basis of Human Behavior*," en A.P. Weiss (1925/1994).
von WRIGHT, G.H. 1971. *Explanation and Understanding*. Ithaca, NY.: Cornell University Press.
WRIGHT, L. 1973. "Functions." *Philosophical Review* 82: 70-86.
——. 1976. *Teleological Explanations: an Etiological Analysis of Goals and Functions.* Berkeley: University of California Press.
WRIGHT, R. 2000. *The Moral Animal: The New Science of Evolutionary Psychology*. Nueva York, NY.: Pantheon.

WYCKOFF, L.B. 1951. *The Role of Observing Responses in Discrimination Learning*. Tesis doctoral inédita, *apud* Escobar y Lattal 2011. Bloomington: Universidad de Indiana.
——. 1954. "A Mathematical Model and an Electronic Model for Learning." *Psychological Review* 61: 89-97.
YALOWITZ, S. 2005. "Anomalous Monism," en E.N. Zalta, ed. 2008.
YATES, F.A. 1966. *El Arte de la Memoria*. Traducción de I. Gómez de Liaño. Madrid: Siruela, 2005.
YELA, M. 1974. *La estructura de la conducta. Estímulo, situación y conciencia*. Madrid: Real Academia de Ciencias Morales y Políticas.
——. 1980/1996. "La evolución del conductismo." *Análisis y Modificación de Conducta* 6: 147-180 / *Psicothema* 8: 165-186.
YOVITTS, A., JACOBI, G.T. y GOLDSTEIN, G.D., eds. 1962. *Self-Organizing Systems*. Nueva York, NY.: Spartan.
ZALTA, E.N., ed. 2008. *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*. URL = <http://plato.stanford.edu/, 1 de agosto de 2010.
ZANGWILL, N. 1992. "Variable Reduction Not Proven." *Philosophical Quarterly* 42: 214-218.
ZIMMERMAN, D.W., ed. 2006. *Oxford Studies in Metaphysics II*. Oxford: Oxford University Press.
ZURIFF, G. 1985. *Behaviorism: A Conceptual Reconstruction*. Nueva York, NY.: Columbia University Press.