UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE FILOSOFÍA

INTELIGENCIA ARTIFICIAL : CONDICIONES DE POSIBILIDAD TÉCNICAS Y SOCIALES PARA LA CREACIÓN DE MÁQUINAS PENSANTES

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR

Manuel Carabantes López

Bajo la dirección del doctor

Emilio García García

Madrid, 2014

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOSOFÍA

PROGRAMA DE DOCTORADO EN FILOSOFÍA

INTELIGENCIA ARTIFICIAL

Condiciones de posibilidad técnicas y sociales para la creación de máquinas pensantes

Tesis presentada para la obtención del grado de doctor por la Universidad Complutense de Madrid

2013

Doctorando: MANUEL CARABANTES LÓPEZ

Director: Dr. Emilio García García

Nuestro riesgo no es la aparición de computadoras superinteligentes, sino de seres humanos subinteligentes.

HUBERT DREYFUS

Índice

Resumen

Desde la aparición de las primeras computadoras electrónicas en la primera mitad del siglo XX, la posibilidad de utilizarlas para replicar la inteligencia humana ha sido una de sus aplicaciones más investigadas. Durante más de medio siglo, varias de las más prestigiosas universidades del mundo y las agencias de defensa de los países desarrollados han invertido grandes cantidades de recursos para emplear a algunas de las mentes más brillantes de diversas disciplinas, principalmente de la ingeniería informática, en la tarea de construir una inteligencia artificial (IA) fuerte, es decir, indistinguible de la de un ser humano. Sin embargo, no lo han logrado.

El presente estudio está dedicado a examinar dos cuestiones: las condiciones de posibilidad técnicas y las condiciones de posibilidad sociales de la IA. Las primeras refieren a la mera posibilidad empírica de utilizar las computadoras electrónicas como soporte para la replicación de la inteligencia. Hay quienes opinan que el fracaso histórico de la IA es síntoma inequívoco de que se trata de un proyecto técnicamente imposible. Por nuestra parte distinguiremos entre dos programas de investigación de la IA: el simbólico y el subsimbólico. El simbólico es el que pretende replicar los procesos mentales, mientras que el subsimbólico aspira a replicar el funcionamiento del cerebro. Nuestra conclusión será que, desde el punto de vista de la psicología comprensiva y del pragmatismo más elemental, la IA simbólica es, en efecto, técnicamente imposible. En cambio, argumentaremos que, desde el punto de vista de la neurociencia, no hay ninguna imposibilidad técnica de principio en el enfoque fisicalista de la IA subsimbólica. Aplicaremos, por tanto, metodologías distintas para sendos análisis, una pluralidad que viene impuesta por las diferencias irreductibles entre ambos objetos de estudio: la mente y el cerebro; el pensamiento y la materia.

En cuanto a las condiciones de posibilidad sociales de la IA, refieren a la adecuación de esta tecnología a los intereses de la sociedad que ha de financiarla y asimilarla a su modo de vida. El dato histórico antes mencionado de que durante más de medio siglo se han financiado multitud de proyectos destinados a construir una IA incluso a pesar del persistente fracaso de los investigadores es una prueba concluyente de que la IA es una tecnología que interesa a los grupos de poder que administran la riqueza de la sociedad industrial avanzada en la que vivimos. Desvelaremos cuáles son esos intereses mediante la teoría social de la primera escuela de Frankfurt, la de Horkheimer, Adorno y Marcuse. Nuestra metodología en este asunto será, pues, la de la dialéctica negativa del marxismo crítico de estos tres filósofos.

Para llevar a cabo ambos análisis, de las condiciones de posibilidad técnicas y sociales de la IA, será menester un extenso estudio preliminar dedicado a exponer los elementos requeridos por la metodología. Así, en el capítulo primero identificaremos a los agentes sociales que han definido la IA en el imaginario popular. En el segundo rastrearemos los orígenes mitológicos y técnicos de los intentos por replicar la mente humana. En el tercero describiremos las características formales, materiales y pedagógicas de las computadoras electrónicas. En el cuarto comenzaremos por exponer varios conceptos de filosofía de la ciencia, seguidos de sendas exposiciones del paradigma cognitivista de la psicología y de la neurociencia, por ser ambas las ciencias en las que se basan respectivamente la IA simbólica y la IA subsimbólica. En el quinto aumentaremos el nivel de concreción de lo visto en el capítulo anterior, pues examinaremos una teoría cognitivista de la inteligencia en la mente y una teoría fisicalista de la inteligencia en el cerebro, junto a las cuales propondremos una teoría holista de la inteligencia como referente nuestro. El último capítulo preparatorio será el sexto, dedicado a la Historia de la IA. Y ya, finalmente, en el séptimo abordaremos las condiciones de posibilidad técnicas de la IA, y en el octavo, las sociales.

Abstract

Since the advent of the first electronic computers in the first half of the twentieth century, the possibility of using them to replicate human intelligence has been one of its most pursued and investigated uses. For over half a century, defense agencies of the developed countries and many of the most prestigious universities in the world have invested large amounts of resources to employ some of the most brilliant minds from various disciplines, mainly of computer engineering, in the task of constructing a strong artificial intelligence (AI), i.e. indistinguishable from that of a human. However, they have not succeeded at all.

The present study is devoted to examining two issues: the technical and the social conditions of possibility of AI. The first refer to the mere empirical possibility of using computers as support for electronic intelligence replication. Some argue that the historical failure of AI is unequivocal sign that it is a technically impossible project. For our part we will distinguish between two research programs in AI: the symbolic and subsymbolic. The symbolic one (also known as GOFAI after "good old-fashioned AI", or "language-of-thought AI" after Jerry Fodor's theory) is that which attempts to replicate the mental processes, whereas the subsymbolic one aims to replicate the brain function. Our conclusion will be that, from the point of view of descriptive psychology and the most fundamental common sense pragmatism, symbolic AI is indeed technically impossible. However, we will argue that, from the point of view of neuroscience, there is no technical impossibility of principle on the physicalist approach of subsymbolic AI. Therefore, we will apply two different methodologies for analysis, a diversity that is imposed by the irreducible differences between the two objects of study: the mind and brain, thought and matter.

As for the social conditions of possibility of AI, they refer to the suitability of this technology to the interests of the society that has to finance and assimilate it to its way of life. The aforementioned historical fact that for over half a century many projects to build an AI have been funded even despite the persistent failure of investigators is conclusive evidence that AI is a technology that interests the power groups who manage the wealth of the advanced industrial society in which we live. We will reveal what those interests are through the social theory of the first Frankfurt School of Horkheimer, Adorno and Marcuse. Our approach in this matter will therefore be the negative dialectics of critical Marxism of these three philosophers.

To carry out both tests, the technical and social conditions of possibility of AI, will require an extensive preliminary study to describe the elements required by the methodology. Thus, in the first chapter will identify social agents who have helped to define the AI in the popular imaginary. The second will trace the mythological and technical origins of the attempts to replicate the human mind. In the third will describe the formal, material and pedagogical features of electronic computers. In the fourth will start for exposing various concepts of philosophy of science, followed by descriptions of the cognitive paradigm of psychology and neuroscience, being both the sciences in which respectively symbolic and subsymbolic AI are based. In the fifth will increase the level of detail seen in the previous chapter, as will examine a cognitivist theory of intelligence in the mind and a physicalist theory of intelligence in the brain, besides proposing a holistic theory of intelligence as ours. The sixth chapter will be the last preparatory one, dedicated to the history of AI. And finally, in the seventh will board the technical conditions of possibility of AI, and in the eighth, the social ones.

Extended summary

Artificial intelligence (AI) may seem at the outset as a topic of science fiction that has nothing to do with philosophy. However science fiction, when valuable, raises important philosophical questions. In particular, artificial intelligence has been treated by the best novelists of the genre, such as Isaac Asimov, Philip K. Dick and William Gibson, as an excuse to reflect on many great philosophical questions, of which here we highlight two: what is to be human and the dystopian consequences of the free of the instrumental or subjective reason from the ethical imperatives dictated by the axiological or objective reason. These are the two main themes that will be discussed in this study. The first, through the examination of the technical conditions of possibility of AI. The second, by doing so with its social conditions of possibility.

All technologies, understood in a broad sense that covers both the technique and the instruments produced by it, must satisfy two types of conditions of possibility, i.e. requirements to be possible. On one side are the purely technical conditions, which relate to the empirical possibility to manipulate nature in order to obtain the technology in question. Thus, for example, electronic computers are technically possible, as they are in effect, and all which is in effect is at least possible. However, there are social groups, such as the Amish, that exclude them from their ordinary lives because they do not suit their interests, governed by reasons of various kinds. Such rejection points straight to other kind of conditions of possibility that every technology must meet: the social ones, referring to its compatibility with the spirit, i.e. with humanity that transcends mere nature. Nature and spirit, therefore, impose their respective demands to every technology to become real.

The technical conditions of possibility of AI will be examined in the seventh chapter, and the social ones in the eighth. But first it is necessary to do an extensive preliminary work that will occupy the remaining chapters: from first to sixth. We will begin by describing the vulgar notion of AI, which is being handled at street level, and is roughly that of a machine with intellectual skills similar to those of a human. Strictly speaking, considering the meaning of the acronym "AI", a machine that imitated the thought of a lower animal would also be an artificial intelligence. However, the subject of this study will not be the conditions of possibility of duplication of any intelligence, but only of human intelligence, which matches with the ordinary notion of AI. Thus defined in anthropocentric terms, the truth is that AI is a technology that does not yet exist. There are machines that perform amazing intellectual tasks, like playing chess better than the best human chess player, but they cannot be considered true artificial intelligences. Their inferiority to us is mainly because the lack of two skills: in the social world, they lack of the natural language, and in the physical world, they do not have our versatility to make a passable attempt at almost anything.

AI is, therefore, a technology that does not yet exist and hence its vulgar notion cannot proceed from the daily contact with it. In our view, its presence in the popular imaginary is a result of the action of three agents: marketing, AI researchers and science fiction. Marketing has been proclaiming for the last decades that we can purchase smart material goods. TVs, phones, cars and even non-mechanical objects such as tissues: intelligence is profuse in all of them, and is attributed in proportional degree to the object's ability to understand and effectively meet the orders or wishes of its owner. AI researchers, meanwhile, have long been promising that are close to replicating human intelligence. Since the founding of their discipline, back in the 50s, have continued to make predictions that never met, and not just predictions, but statements against evidence that their machines are able to do things that simply do not. His strategy, in many cases, is to repeat the same lie over and over again with the hope that listeners finish believing it. Many causes can be applied to explain such dishonest conduct, but the key one is that they behave so to get funding. They are

scientists, but first of all they are human beings with professional ambitions and material needs to cover. Finally, with regard to science fiction, it is perhaps, in feedback to the other two, the main agent that has shaped the vulgar notion of AI. In the worst cases, through literary and audiovisual works made for the sole purpose of collecting benefits, and in the best ones, with works that have used the theme, as we said, to address philosophical issues.

However, the success of AI as commercial hook and as justification for items of public budgets for research must have a cause. This issue, the source of attraction for the thinking machines, will be addressed in the *second chapter* in a double sense: anthropological and historical. The origin in the anthropological sense will be found out by examining the meaning of the myths of the recreation of man by man. Led by the specialist in cybernetics philosopher André Robinet, we will appreciate the similarities and differences of these myths in monotheistic and polytheistic religions. On the other hand, will trace the historical origins of the first automata of thought. It was in the seventeenth century, at the dawn of Modernity, when, for technical and social reasons, could be undertaken the first plausible attempts to build thinking machines. Specifically, robots capable of performing mathematical calculations. Among those who went on the adventure of trying to build these devices there are illustrious names such as Pascal and Leibniz. However, the first to achieve it was a humble anonymous mathematician, the German Wilhelm Shickard in 1623. Since then, it took two hundred years until an English inventor, Charles Babbage, carried the matter a step further. With its Analytical Engine, Babbage raised the design of a machine almost as versatile as current electronic computers. Unfortunately, he only had time and money enough to build half of it.

Returning to the historical origin of the thinking machines, in the seventeenth century the possibility of building such devices, able to perform mental operations, or at least imitate the behavior produced by them, was received unevenly throughout the philosophy of the time. In the latter part of the second chapter we will compare the different positions of Descartes and Hobbes. The first one, often claimed as the father

of modern rationalism, based on his metaphysical dualism of substances and his Christian convictions to argue against the technical possibility to perfectly replicate human thought. In his view, there were two faculties of our intellect that could not be produced by mechanical combinations, and they are just the two mentioned in the previous chapter: in the social world, the language, and in the physical world, the flexibility to undertake any task. Descartes thus showed exceptional acuity, when passing over three hundred years the diagnosis of the two biggest obstacles to AI.

Instead, Hobbes, empiricist and father of modern mechanistic psychology, not only did not appreciate any technical hurdle in the venture of the automata of thought, but quite the opposite: if they were built, they would have confirmed his materialistic thesis. Let us note also that the choice of these two authors, Descartes and Hobbes, as exponents of the philosophical context in which emerged the first automata of thought, is not only because their respective representativeness within the rationalist and empiricist branches, but also because Cartesian substance dualism persists today to some extent in a stream of AI called symbolic AI, while Hobbes corporalism is present in the other great approach of this discipline: subsymbolic AI.

The *third chapter* will be devoted to describe the traits of computers, which are essential to elucidate the technical conditions of possibility of AI as they are the machines used by the scientists in their pursue to achieve it. We will distinguish three dimensions in them: formal, material and pedagogical. At a formal level electronic computers are formal systems, i.e. sets of symbols over which rules apply to form and transform expressions. To illustrate the potential of formal systems, and any limitations, we will describe the operation of Turing machines, ideal artifacts that do exactly the same as any real computer: executing algorithms, which can be defined as finite sets of instructions whose execution yields a desired result. Reviewing the history of mathematical automata made in the previous chapter will receive a new meaning when we discover that Alan Turing conceived his machines with the intention to formally define the computing tasks that until that moment were carried out by networks of human computers endowed with not a particularly bright intelligence.

Regarding the material dimension of electronic computers, we will talk on the sum of components and miniaturization. The latter is a technique that for decades has improved the performance of these machines exponentially but that, however, is about to run into insurmountable physical limits. And finally, will reveal the pedagogical condition inherent to all technique in general, and specifically how the pedagogics of computers affects to what can be done with them. In a text as old as Plato's *Phaedrus* we will find constraints that weigh decisively on the software design and cannot be overcome even with the latest technology.

As already mentioned, AI is divided into two main branches or research programs: symbolic and subsymbolic AI. The first pursues to replicate the mind, and the second, the brain. Now, mind and brain can be designed in many ways. To detail the models of mind and brain that claim to be replicated, respectively, by symbolic and subsymbolic AI, will be the goal of *chapter four.* To do this we will begin dedicating a section to expose a number of concepts in philosophy of science that will be needed, such as Thomas Kuhn's paradigm and Imre Lakatos' research program, and to address three distinctions: realism and instrumentalism, rationalism and relativism, explaining and describing. In the first we will show that science, always in its modern sense, is a mere instrument for the domination of nature that cannot hold truth claims. In the second will argue in favor of the thesis that the scientific method, understood as an algorithmic procedure that ensures obtaining knowledge, is a myth. Instead, the real scientific activity proceeds by applying logical and psychological strategies in both the contexts of discovery and justification. The existence of psychological strategies in the context of justification is evidence for the relativistic approach. And finally, the third distinction will be between explaining and describing. Natural sciences employ an explanatory method, while social sciences use a descriptive method.

The explanatory method is based on a molecular bottom-up approach, that tries to reduce complex phenomena to simple or atomic facts. Instead, the descriptive method has a molar approach that goes top-down to catch the meaning of simple phenomena in the total context in which they occur. The difference between both

affects in a particular way to psychology, since, as it studies the mind, and as the mind is a product of the interaction of natural and social factors, psychology is forced to integrate explanation an description, molecularity and molarity. Given the impossibility of such a synthesis, the various paradigms of psychology have chosen to favor one of the two methods, which leads to the distinction between explicative psychology (Physiologische Psychologie) and descriptive psychology (Geisteswissenchaftliche Psychologie). In the case of cognitive psychology, or cognitivism, which is the paradigm that supplies its model of the mind to the symbolic AI, the choice is the explanatory approach, characteristic of natural sciences.

After the digression on concepts of philosophy of science, we will discuss the essential features of cognitivism, which are five. Two are alleged nuclear, and the other three are methodological. The nuclear assumptions are the internalist and the information processing thesis. The first one holds, as we anticipated at the end of the second chapter, a kind of Cartesian substance dualism. Not ontologically, i.e. it does not state in the space age that mind and body are distinct substances, but it does methodologically as it holds that for explaining the operation of an intentional system, as the human mind is, it is necessary to postulate the existence of a mental level of representation independent of the biological processes from which it arises, namely: that the mind is explainable independently of the body. As for the information processing thesis, also known as the computational metaphor, it characterizes the mind as an information processor similar to an electronic computer. The consequence of this postulate is a circular unproductive relationship, devoid of tension, between cognitive psychology and symbolic AI: the first assumes that the mind is like a computer, and the second aims to use computers to replicate the functioning of the mind. In this circle AI researchers have moved for decades, and from it they have taken their previously commented reckless confidence.

Having described the other features of cognitivism, we will characterize the science that provides the brain model to the subsymbolic AI. This is the neuroscience. Unlike psychology, which is abundant in opposing schools, neuroscience has a long

history of accumulative progress around a single paradigm recognized by the scientific community: the modularity. However, it was not always so. Before reaching consensus on modularity, neuroscience discussed for centuries between the approaches of holism and localizationism. Modularity, as we shall see, holds a certain localizationism, but moderate and conciliatory with holism, while holding the precise location of only very basic functions, instead of the complex ones that Joseph Gall believed he had precisely located in the late eighteenth century.

The fourth chapter will conclude contrasting cognitivism and the materialist reductionism. Both are eliminativist approaches: the first is intended to explain the mind independently of the body, and the second to explain behavior without regard to the mind. Faced with any kind of eliminativism, our position is that of emergentism, i.e. the theory that the mind emerges from the brain not acting on it, but as a mere epiphenomenon. Yet despite the difficulties in the study of the mind because of the mentioned requirement to synthesize explanatory and descriptive approaches by the biosocial constitution of mind, it is an essential task in our view, while materialistic reductionism, as we will demonstrate by arguments borrowed from Hilary Putnam, falls into the error of assuming the transitivity of explanations.

Once described the general characteristics of the paradigms of mind and brain on which symbolic and subsymbolic AI are respectively based, in the *fifth chapter* will increase the level of detail. We will expose a model of intelligence in the brain by a computer engineer, Jeff Hawkins, and a cognitivist model of intelligence at the mental level developed by Roger Schank, a renowned computer AI researcher who in recent times has engaged to psychology. Schank's cognitivist model of intelligence will help us to appreciate from within the irresolvable difficulties with which symbolic AI stumbles in its attempt to use formal systems, which is what in the background all electronic computers are, to replicate the distinctive versatility of human intelligence. If the epistemological-logical psychologism, such as Stuart Mill's, that Husserl fought aimed to reduce logic to psychology, cognitivism claims to do just the opposite: to reduce psychology to formal logic systems.

Towards the end of the exposure of Schank's cognitivist model of intelligence, we will chart a common thread with the subject of the myth of the scientific method discussed in the previous chapter. Our thesis is that what underlies symbolic AI and the myth of the method in the above specified sense is the same positivist claim: finding algorithms for generating theories, whether scientific theories in the case of the scientific method or pre-scientific ones in the case of symbolic AI. In both cases the goal is to automate the production of knowledge, an impossible task that, however, is a sign of our times, and is in continuity with the purposes for which they were created the human calculators networks and later the electronic computers.

In the fifth chapter we will also discuss the theory of multiple intelligences (MI theory) of Howard Gardner, which, unlike the theories of intelligence by Schank and Hawkins, is the most successful in our view for two main reasons. The first is its relativistic characterization, in the best sense, as far as it considers intelligence as a faculty which is multiple, distributed and contextualized, just the opposite to the unitary, solipsistic and ethnocentric approaches that predominate in psychometrics. The other reason for our adherence to the MI theory is that the criteria proposed by Gardner for identifying an intelligence cover many different factors, from biological to cultural and historical ones. Such amplitude reflects the complex nature of intelligence, rather than avoiding it like those psychologists who define intelligence in operational terms as that faculty which is measured by intelligence tests.

Of course, in a philosophical reflection on AI and on the nature of intelligence could not miss an examination of the Turing test. That will be the issue of the last section of the fifth chapter. We will expose it, will expose the refutation against it by John Searle with his famous Chinese room argument, will discuss the objections that have been raised by Searle's argument and collected by himself, will discuss Searle's responses to those objections, and finally will expose our own objections to those answers. We will dive therefore into a multi-layered dialogue that culminates with a defense of the Turing test's behavioral criteria of intelligence. Also will evaluate the importance that the English mathematician, like Descartes, gave to the language.

In the *sixth chapter* we will review the history of AI since its official founding at the Dartmouth conference, held in 1956, to the present. At its birth the AI was assigned to the cognitivist paradigm, officially founded at a MIT's symposium held in the same year a month later, and also involving two of the attendees who were at Dartmouth: Allen Newell and Herbert Simon, the authors of the first AI. Or, more correctly, the first attempt of an AI, as the Logic Theorist, which was the name of that computer program, was not a true AI. In fact, as we will see throughout the chapter, this discipline has never been able to create an AI in the strong sense that matches the vulgar notion of AI. Newell and Simon, along with other prominent researchers such as Marvin Minsky, gripped from the beginning to the research program of symbolic AI, which is the duplication of mind understood in cognitivist terms, while they used their power to discredit those like Frank Rosenblatt who dared to venture into the subsymbolic AI approach, which is, remember, the duplication of neural networks or, at least, the simulation of its operation. The situation changed towards the 80s, and since then both points of view, symbolic and subsymbolic, coexist and even complement each other. However, none has built a machine as intelligent as a human being. Each research program has its own problems, which will be examined in this chapter. The main one of the symbolic AI is the domain limitation, i.e. the lack of versatility, which is, as we have noted, one of the two distinctive characteristics of our intellect. The other one, natural language, has also not been replicated yet.

Building on the theme of the history of AI, we will also discuss some of the strategies, techniques and most relevant architectures invented by researchers in their attempt to create thinking machines. This exhibition will lead to the presentation of opposing alternatives such as the usual symbolic vs subsymbolic AI, human vs alien AI, strong vs weak AI, abstract vs situated AI, and strong vs weak methods. These are concepts that, although coined by engineers and mathematicians, also refer to philosophical problems. For example, the second AI built by Newell and Simon was the GPS, which stands for General Problem Solver, a program designed as its name suggests to solve any problem. To achieve such a feature its creators opted for a weak

method, that is, a heuristic method of finding solutions that did not depend on any prior knowledge, facing strong methods, which are those that do depend on a database that provides useful information to solve problems. The Cartesian method for directing well the reason and finding out truth in science was a weak method, as the French philosopher renounced in principle to rely on memory, which is the human equivalent of databases in electronic computers. The history of AI, as shown by this example, has come in just over half a century many of the theories of knowledge that have been proposed by philosophers over 2,500 years.

Within the AI there has been researchers who have gone beyond the art and have taken philosophy to identify the epistemological root causes of certain problems. In this regard will see what John McCarthy and Patrick Hayes noted in his 1969 article *Some philosophical problems from the standpoint of artificial intelligence.* Of all the problems mentioned in the text, two are the most relevant: the frame problem and the qualification problem. The frame problem concerns the issue of updating a heavily interconnected database, so that the modification of only one data can implement a butterfly effect that results in the need to modify many others. The problem of qualification, however, is less technical and more philosophical. By its nature be such, it is the intractable problem we will use preferably in the next chapter, the seventh, to refute the technical possibility of symbolic AI. The qualification problem refers to the inability to develop an explicit list of the conditions of validity of the rules that humans use to cope with the world, both in its physical and social dimensions. For example, we operate with the standard "ships are used to navigate", but it only applies if a lot of impossible-to-explicit conditions are met, such as that the wreck must not have a hole in the hull. In turn, the size of the hole is a condition subject to conditions such as its size, since a small hole might not be an impediment to the navigability. Thus, the rules have validity conditions which in turn may refer to others in indefinite sequence.

With all the research work carried out in the six previous chapters, in the *seventh chapter* we already will be able to address one of the two great questions of this study: the technical conditions of possibility of AI. As our field is philosophy, we

will examine such technical conditions of possibility not from the point of view of engineering, but of epistemology. The more interesting exam will be the one on the conditions of possibility of symbolic AI. However, in the subsymbolic AI will also discuss philosophical issues of relevance, such as the relationship between mind and body.

To analyze the technical conditions of possibility of AI we will employ a dual approach, methodological pluralism that is imposed by the ontological difference between symbolic AI, which, as already noted, has the goal to replicate the mind as conceived by the cognitivist paradigm of psychology, and subsymbolic AI, which aims to duplicate the brain. On symbolic AI we will discuss and expand the critique made by Hubert Dreyfus, a philosopher who uses a large battery of arguments against the cognitivist model of the mind. Dreyfus, with the occasional collaboration of his brother Stuart, became famous in the 60s, just when the AI as a discipline had become, for a series of attacks that earned him the hate of all researchers in this field.

Dreyfus's arguments owe their strength on a phenomenological conception of mind, i.e. molar, against the molecular one sustained by symbolic AI and cognitivism. Dreyfus takes its molar approach from existentialist philosophers like Heidegger and Merleau-Ponty and Gestalt psychologists like Max Wertheimer and Kurt Goldstein. With these tools, he does a double work of destruction of symbolic AI and cognitivism. His first step is to identify four cognitive processes unexplainable from a molecular approach of mind. These are the fringes of consciousness, ambiguity tolerance, essential and inessential discrimination and perspicuous grouping. Fringes of consciousness, rather than a process, is the place located beyond the consciousness in which occur those cognitive processes that we are not aware of and have no control on. That's where the chess master sees the key play on the board without having to calculate hundreds of possible move sequences as computers do, a skill that is possible thanks to the gestaltist notion of figure and ground, as the history of games played acts as a background that determines what is presented in the consciousness of the master as a figure that attracts his attention. The other three non-replicable cognitive processes by symbolic AI they all take place in the fringes of consciousness. Thus, we

have no awareness of how we tolerate ambiguity, discriminate between essential and inessential characteristics and apply perspicuous grouping, i.e. the skill on which pattern recognition is based. With this review Dreyfus demonstrates the impossibility in principle of symbolic AI, but only in the aspect of human AI, which is aimed at replicating the human mind in a realistic way.

So it would remain the possibility of alien AI, defined as the approach that aims to build an artificial mind whose behavior becomes indistinguishable from the human one even if it does through different cognitive processes. To prove also the impossibility of alien AI, Dreyfus launches a second wave of destruction on symbolic AI and cognitivism: investigates the four assumptions underlying symbolic AI in general. These are, from least to most important: biological, psychological, epistemological and ontological. The biological one holds that the brain is a discrete state machine similar to an electronic computer. The psychological, also known as the strong physical symbol system hypothesis, postulates that the human mind uses computational processes to produce intelligent behavior, which is just the same as holding one of the two nuclear assumptions of cognitivism: the computational metaphor. The epistemological, also known as the physical symbol system hypothesis is silent on how the human mind works, but by acquiring a lower degree of commitment, states only that a formal system is sufficient to produce intelligent behavior so that all intelligent behavior can be formalized. And finally, the ontological assumption is, in terms of Lakatos' theory of science, the core of the subsymbolic AI research program, and thus refuting it the whole symbolic AI falls down, including the alien branch. The ontological assumption is simply another name for the logical atomism of Bertrand Russell and the early Wittgenstein: the assumption that the world can be represented as a structured set of descriptions which are built up from primitive or atomic expressions.

Dreyfus' demolition work is progressive. Begins by removing the biological assumption and ends up refuting the ontological one. For our part, we will extend Dreyfus' critique in two ways. First, accumulating evidence taken from other sources to provide even stronger support to his arguments. This will be done by referring to

matters discussed in previous chapters, and not by chance, but precisely in order to make them converge here. Thus, for example, we will appeal to the exposure of the elements of neuroscience in the fourth chapter to enhance the refutation of the biological assumption, and will rescue the circular structure of understanding outlined by Heidegger and Gadamer as seen in the fifth chapter to argue against the objectivist epistemology of logical atomism.

The second way in which we will extend Dreyfus' Critique is by a section that will face the pragmatist arguments of the latter Wittgenstein against atomistic arguments of *Tractatus*' Wittgenstein. Since symbolic AI is based on the ontology of logical atomism, who better than Wittgenstein himself, the greatest exponent of the doctrine, to fight it. We will rescue the brightest moments of the *Blue and brown books* and the *Philosophical investigations* where the German philosopher attacks his former beliefs. In this respect the model of the mind of Roger Schank, exposed in the fifth chapter and revisited here, will be enlightening, since it is a cognitivist model of mind, based therefore on the manipulation of symbolic expressions, but at the same time is strongly influenced by pragmatism, as seen in his frequent allusions to John Dewey and his attempt to explain the practical skills of ordinary life. Because Schank is not trained in philosophy, but only in AI and psychology, hes does not realize that is trying to merge two incompatible ontologies: the logical atomism underlying the cognitivism on one hand and the pragmatism on the other.

Cognitivism is based on the assumption that the mind is like an electronic computer, and as such it has to operate running unambiguous rules at all times. However, as Wittgenstein noted, the mind does not always operate following rules, but simply some things are as they are, and we cope with them without been capable of expliciting our behaviour in an algorithmical way, and even if we could, such a formulation would not be enough to perform. The qualification problem pointed by McCarthy and Hayes, to which we referred earlier, is of high relevance on this issue, since it states that many rules rest over conditions of validity which are impossible to explicit, althought such explicitation is imperative for computers.

About subsymbolic AI, we share with Dreyfus the assessment that it is technically possible if the approach adopted is situated AI, that is, against abstract AI, the one wich aims to create thinking machines by placing them in a body. However, instead of arguing from the philosophy of Merleau-Ponty as Dreyfus does, we will do it from neuroscience. As neuroscientist Antonio Damasio says, believing in the possibility of replicating the human brain without a body is a kind of materialist Cartesianism. To evaluate the proximity in time of the construction of a subsymbolic AI we will examine the latest techniques for modeling networks of neurons, such as artificial evolution. The biggest challenge is to decipher the subsymbolic AI starting connectome of the human brain, i.e. the initial wiring of neural networks determined by genetic factors that is modified by the experience to model the intelligent adult brain.

On the *eighth chapter*, the last one, the theme will be the social conditions of possibility of AI. To resolve whether such technology has a place in today's society, and to predict its most plausible uses, we must begin with an analysis of society itself. We will do it from the point of view of the Frankfurt School, and more specifically, the theories of Horkheimer, Adorno and Marcuse on advanced industrial society and mass culture. The Enlightenment, understood as the historical process of rationalization to free man from fear and to emancipate, has paradoxically resulted, as Horkheimer and Adorno said, in a relapse in terror and slavery. The subjective or instrumental reason, which is the creator of modern science and technology, has been freed from objective or axiological reason, which is the rector of human actions. Our contribution to this philosophy of history will consist mainly on plotting the psychological profile of the individual resulting from it. This profile is characterized, in our view, because of its similarity with two diseases of the mind: psychopathy and schizophrenia.

The psychopath is the winner, successful man of our time, a time that, as measured in purely instrumental terms, its only rationality criterion is success. To describe the features of psychopathy we will use the theory of psychologist Robert Hare, considered the world's foremost expert on the subject. As Horkheimer and Adorno used in their *Dialectic of Enlightenment* the characters of the Marquis de Sade

to portray the individual guided by subjective reason alone, we will use the description of the psychopath according to Dr. Hare, obtained by experimental psychology and cognitive psychophysiology. At least from a philosophical point of view, and perhaps also of clinical diagnosis, the four libertines of *120 days of Sodom* have the profile of a psychopath: egocentric, without a trace of guilt, lack of empathy, manipulative, cold, impulsive, addicted to extreme experiences. However, they do not show themselves that way all the time, because they know that if they did, they could not carry out their plans successfully. In practice, such behaviors operate in society, but are formally rejected by a completely opposite set of values that, under such opposition, generate schizophrenia. While one thing is done, the opposite is said. In Sade*'s Misfortunes of Virtue*, naive Justine was brutally exploited by characters that in public gave the appearance of virtue that society demands. A doctrine this, the double-truth, which generates schizophrenia, and is essential for success in today's society.

AI, while an automating technology and therefore one that could reduce the socially necessary and individually repressive labor time, contains a revolutionary power to transcend the current state of things. However, that revolution will not happen necessarily as positivism states appealing to alleged laws of social progress, but from the Marxist and antimechanistic perspective held by the Frankfurt School it depends on the action of a historical subject that should be conscious of the possibility of such becoming from the point of view of subjective reason and its moral necessity from the point of view of the objective reason. The making of that consciousness is, however, impossible due to the global totalitarian action of mass culture, as demonstrated by Horkheimer and Adorno. As long as the cultural industry operates the closing of the universe of speech and action with the efficacy exhibited so far, the progress of history will remain suspended even though technologies emerge with a revolutionary potential as high as the AI.

Introducción

La inteligencia artificial (IA) puede parecer, de entrada, un tema de ciencia ficción que nada tiene que ver con la filosofía. Sin embargo, la ciencia ficción, cuando es valiosa, plantea grandes cuestiones filosóficas. En concreto, la inteligencia artificial ha sido tratada por los mejores novelistas del género, tales como Isaac Asimov, Philip K. Dick y William Gibson, como una excusa para reflexionar sobre muchas grandes cuestiones filosóficas, de las que aquí destacamos dos: qué es ser humano y las consecuencias distópicas de la liberación de la razón instrumental o subjetiva respecto de los imperativos éticos dictados por la razón axiológica u objetiva. Estos son los dos temas centrales de los que nos ocuparemos en el presente estudio. El primero lo abordaremos a través del examen de las condiciones de posibilidad técnicas de la IA. El segundo, haciendo lo mismo con sus condiciones de posibilidad sociales.

Toda tecnología, entendiendo ésta en un sentido amplio que abarca tanto la técnica como los instrumentos producidos por ella, ha de satisfacer dos tipos de condiciones de posibilidad, es decir, de requisitos necesarios para ser posible. Por un lado están las condiciones puramente técnicas, que refieren a la posibilidad material de manipular la naturaleza para obtener la tecnología en cuestión. Así, por ejemplo, las computadoras electrónicas son técnicamente posibles, pues existen en efecto, y todo lo efectivo es, cuando menos, posible. Sin embargo, hay grupos sociales, como los Amish, que las excluyen de su vida ordinaria porque no se adecúan a sus intereses, regidos éstos por causas de diversa índole. Tal rechazo apunta al otro tipo de condiciones de posibilidad que debe cumplir toda tecnología: las sociales, que refieren a su compatibilidad con el espíritu, es decir, con lo humano que trasciende la mera naturaleza. Naturaleza y espíritu, por tanto, imponen sus respectivas exigencias.

Las condiciones de posibilidad técnicas de la IA las examinaremos en el capítulo séptimo, y las sociales, en el octavo. Pero antes será necesario un extenso trabajo preliminar que nos ocupará el resto de los capítulos: desde el primero al sexto. Empezaremos por describir la noción vulgar de IA, que es la que se maneja a pie de calle, y que a grandes rasgos es la de una máquina con unas destrezas intelectuales similares a las de un ser humano. En sentido estricto, atendiendo al significado de las siglas "IA", una máquina que imitase el pensamiento de un animal inferior también sería una inteligencia artificial. Sin embargo, el tema de este estudio no serán las condiciones de posibilidad de la duplicación de cualquier inteligencia, sino sólo de la inteligencia humana, que es la que coincide con la noción vulgar de IA. Así definida, en términos antropocéntricos, lo cierto es que la IA es una tecnología que todavía no existe. Hay máquinas que hacen cosas asombrosas, como jugar al ajedrez mejor que el mejor ajedrecista humano, pero no se puede decir, en rigor, que sean auténticas inteligencias artificiales. Su inferioridad respecto a nosotros se debe, principalmente, a que carecen de dos habilidades: en el mundo social, del lenguaje, y en el mundo físico, de nuestra versatilidad para hacer un intento pasable en casi cualquier cosa.

La IA es, por tanto, una tecnología que todavía no existe y, en consecuencia, su noción vulgar no puede proceder del contacto mundano con ella. A nuestro juicio, su presencia en el imaginario popular es resultado de la acción de tres agentes: la mercadotecnia, los investigadores de la IA y la ciencia ficción. La mercadotecnia lleva décadas pregonando que podemos adquirir bienes materiales cada vez más inteligentes. Televisores, teléfonos, automóviles e incluso objetos no mecánicos como los tejidos: la inteligencia es un atributo que abunda en todos ellos, y que se predica en un grado proporcional a la capacidad del objeto para comprender o satisfacer eficazmente las órdenes o deseos de su propietario. Los investigadores de la IA, por su parte, llevan mucho tiempo prometiendo que están cerca de replicar la inteligencia humana. Desde la fundación de su disciplina, allá por los años 50, no han cesado de hacer previsiones que jamás se cumplen, y no sólo predicciones, sino de afirmar contra la evidencia que sus máquinas son capaces de hacer cosas que, sencillamente, no

hacen. Su estrategia, en muchos casos, es la de repetir la misma mentira una y otra vez con la esperanza de que los demás terminen creyéndola. Se pueden postular muchas causas de tan deshonesta conducta, pero la fundamental es que se comportan así para conseguir financiación. Por encima de su condición de científicos, son seres humanos con ambiciones profesionales y con necesidades materiales que cubrir. Por último, respecto a la ciencia ficción, es quizás, en retroalimentación con los otros dos, el agente que más ha contribuido a configurar la noción vulgar de IA. En los peores casos, a través de obras literarias y audiovisuales elaboradas con la única finalidad de recoger beneficios, y en lo mejores, con obras que han utilizado el tema, como decíamos al comienzo, para abordar asuntos filosóficos.

Ahora bien, el éxito de la IA como gancho comercial y como argumento para justificar partidas de los presupuestos públicos destinados a investigación debe tener una causa. Este asunto, el origen de la atracción por las máquinas pensantes, lo abordaremos en el *capítulo segundo* en un doble sentido: antropológico e histórico. El origen en sentido antropológico lo descubriremos examinando el significado de los mitos de la recreación del hombre por el hombre. De la mano del filósofo especialista en cibernética André Robinet apreciaremos similitudes y diferencias de estos mitos en las religiones monoteístas y politeístas. Por el otro lado, rastrearemos el origen histórico de los primeros autómatas del pensamiento. Fue en el siglo XVII, recién iniciada la Modernidad, cuando, por motivos técnicos y sociales, pudieron acometerse los primeros intentos plausibles de construir autómatas del pensamiento. En concreto, autómatas capaces de efectuar cálculos matemáticos. Entre los que se lanzaron a la aventura de intentar construir estos artefactos figuran nombres tan ilustres como los de Pascal y Leibniz. Sin embargo, el primero que lo conseguiría fue un humilde matemático anónimo, el alemán Wilhelm Shickard. Desde entonces, 1623, hubieron de pasar doscientos años hasta que un inventor inglés, Charles Babbage, llevase la cuestión un paso más allá. Con su máquina analítica Babbage planteó el diseño de una máquina casi tan versátil como las computadoras electrónicas actuales. Por desgracia, sólo tuvo tiempo y dinero suficientes para construir la mitad.

Volviendo al origen histórico de las máquinas pensantes, en el siglo XVII la posibilidad de construir semejantes artefactos, capaces de realizar operaciones mentales, o cuando menos de imitar la conducta producida por ellas, fue recibida de manera desigual por la filosofía de la época. En la última parte del segundo capítulo contrastaremos la diferencia de posturas al respecto entre Descartes y Hobbes. El primero, padre del racionalismo moderno, se basó en su metafísica del dualismo de sustancias y en sus convicciones cristianas para argumentar en contra de la posibilidad técnica de replicar de manera perfecta el pensamiento humano. A su entender, había dos facultades de nuestro intelecto que no podían ser producidas por combinaciones mecánicas, y son justo las dos que mencionamos en el capítulo anterior: en el mundo social, el lenguaje, y en el mundo físico, la flexibilidad para acometer casi cualquier tarea. Descartes demostró así una agudeza excepcional, al adelantarse en más de trescientos años al diagnóstico de los dos mayores obstáculos de la IA.

En cambio, Hobbes, empirista y padre de la psicología mecanicista moderna, no sólo no apreciaba ningún impedimento técnico en la empresa de los autómatas del pensamiento, sino que, de haberse logrado construirlos, se habrían confirmado sus tesis materialistas. Apuntemos también que la elección de estos dos autores, Descartes y Hobbes, como exponentes del contexto filosófico en el que surgieron los primeros autómatas del pensamiento, obedece no sólo a su representatividad respectiva dentro de las corrientes racionalista y empirista, sino también a que el dualismo de sustancias cartesiano persiste actualmente, en cierta manera, en una corriente de la IA denominada IA simbólica, mientras que el corporalismo de Hobbes tiene su continuación en la otra gran corriente de esta disciplina: la IA subsimbólica.

En el *capítulo tercero* describiremos las características de las computadoras, imprescindibles para elucidar las condiciones de posibilidad técnicas de la IA en tanto que son las máquinas con las que los científicos pretenden construirla. Lo haremos distinguiendo tres dimensiones en ellas: formal, material y pedagógica. A nivel formal las computadoras electrónicas son sistemas formales, es decir, conjuntos de símbolos sobre los que se aplican reglas para formar y transformar expresiones. Para ilustrar el

potencial de los sistemas formales, y también alguna limitación, describiremos el funcionamiento de las máquinas de Turing, artefactos ideales que hacen exactamente lo mismo que cualquier computadora real, esto es: ejecutar algoritmos, conjuntos finitos de instrucciones rutinarias cuya ejecución arroja un resultado deseado. El repaso de la Historia de los autómatas matemáticos realizado en el capítulo anterior cobrará un nuevo sentido en éste cuando descubramos que Alan Turing concibió sus máquinas con la intención de definir formalmente las tareas de computación que hasta el momento venían siendo efectuadas por redes de computadores humanos dotados de una inteligencia no especialmente brillante. Respecto a la dimensión material de las computadoras electrónicas, hablaremos sobre la suma de componentes y la miniaturización. Esta última es una técnica que durante décadas ha permitido mejorar las prestaciones de estas máquinas de manera exponencial pero que, sin embargo, está a punto de toparse con límites físicos infranqueables. Y, por último, desvelaremos la condición pedagógica inherente a toda técnica en general, y en concreto cómo afecta la pedagogía de las computadoras al alcance de lo que se puede hacer con ellas. En un texto tan antiguo como el *Fedro* de Platón encontraremos limitaciones que pesan decisivamente sobre el diseño de programas informáticos y que no pueden ser superadas ni con la más moderna tecnología.

Como ya hemos mencionado, la IA se divide en dos grandes corrientes o programas de investigación: IA simbólica e IA subsimbólica. La primera pretende replicar la mente, y la segunda, el cerebro. Ahora bien, mente y cerebro puede ser concebidos de muchas maneras posibles. Exponer con detalle de los modelos de la mente y del cerebro que pretenden ser imitados, respectivamente, por la IA simbólica y la IA subsimbólica será el objetivo del *capítulo cuarto*. Para ello comenzaremos dedicando una sección a exponer una serie de conceptos de filosofía de la ciencia que nos harán falta, tales como el de paradigma de Thomas Kuhn y el de programa de investigación de Imre Lakatos, así como para abordar tres distinciones: realismo e instrumentalismo, racionalismo y relativismo, explicar y comprender. En la primera demostraremos que la ciencia, siempre en el sentido de la ciencia moderna, es un

mero instrumento para la dominación de la naturaleza que no puede albergar pretensiones de verdad. En la segunda argumentaremos en favor de la tesis de que el método científico, entendido como procedimiento algorítmico que garantiza la obtención de conocimiento, es un mito. En su lugar, la actividad científica real procede aplicando estrategias lógicas y psicológicas tanto en el contexto de descubrimiento como en el de justificación. La existencia de estrategias psicológicas en el contexto de justificación constituye una prueba en favor del enfoque relativista de la ciencia frente al racionalista. Y, por último, en la tercera distinción abordaremos las diferencias entre explicar y comprender. Las ciencias de la naturaleza emplean un método explicativo, mientras que el de las ciencias sociales es comprensivo.

El método explicativo se fundamenta en un enfoque molecular que va de abajo a arriba, intentando reducir los fenómenos más complejos a fenómenos simples o atómicos. En cambio, el método compresivo tiene un enfoque molar que va de arriba a abajo, para captar el significado de los fenómenos más simples en el contexto total en el que se dan. La diferencia entre ambos métodos afecta de manera singular a la psicología, dado que, por ocuparse ésta del estudio de la mente, y por ser la mente producto de la interacción de factores naturales y sociales, la psicología se ve obligada a intentar integrar la explicación y la comprensión, la molecularidad con la molaridad. Ante la imposibilidad de realizar semejante síntesis, los diversos paradigmas de la psicología han optado por privilegiar uno de los dos métodos. En el caso de la psicología cognitiva, o cognitivismo, que es el paradigma que suministra a la IA simbólica su modelo de la mente, su opción ha sido adoptar el enfoque explicativo, el propio de las ciencias de la naturaleza.

Una vez concluida la digresión sobre conceptos de filosofía de la ciencia, expondremos las características esenciales del cognitivismo, que son cinco. Dos son supuestos nucleares, y las otras tres son rasgos metodológicos. Los supuestos nucleares son la tesis internalista y la tesis del procesamiento de información. La tesis internalista es la que habíamos anticipado al final del capítulo segundo que sostiene el dualismo de sustancias cartesiano. No de manera ontológica, es decir, no afirma en

plena era espacial que la mente y el cuerpo sean sustancias distintas, pero sí de manera metodológica, en tanto que sostiene que para explicar el funcionamiento de un sistema intencional, como es la mente humana, es necesario postular la existencia de un nivel mental de representaciones independiente de los procesos biológicos de los que surge, esto es: que la mente es explicable con independencia del cuerpo. En cuanto a la tesis del procesamiento de información, también conocida como metáfora computacional, caracteriza a la mente como un procesador de información similar a una computadora electrónica. La consecuencia de tal postulado es, como se puede apreciar, una relación de circularidad improductiva, carente de tensión, entre la psicología cognitiva y la IA simbólica: la primera supone que la mente es como una computadora, y la segunda pretende utilizar computadoras para replicar el funcionamiento de la mente. En este círculo se han movido los investigadores de la IA durante décadas, y de él han tomado la confianza optimista que les ha empujado a afirmar de que las máquinas pensantes se conseguirían en poco tiempo.

Una vez descritos los demás rasgos del cognitivismo, pasaremos a caracterizar la ciencia que proporciona a la IA subsimbólica el modelo del cerebro en el que ella se basa. Ésta es la neurociencia. A diferencia de la psicología, en la que abunda la variedad de escuelas contrapuestas, la neurociencia lleva mucho tiempo progresando acumulativamente en torno a un solo paradigma reconocido por toda la comunidad científica: la modularidad. No obstante, no siempre fue así. Antes de alcanzar el consenso sobre la modularidad, la neurociencia se debatió durante siglos entre el holismo y el localizacionismo. La modularidad, como veremos, sostiene un cierto localizacionismo, pero moderado y conciliador con el holismo, en tanto que sostiene la localización precisa sólo de funciones muy elementales, en lugar de las funciones complejas que Joseph Gall creyó haber localizado a finales del siglo XVIII.

El capítulo cuarto lo finalizaremos contrastando el cognitivismo con el reduccionismo materialista. En ambos casos se trata de enfoques eliminativistas: el primero pretende explicar la mente con independencia del cuerpo, y el segundo explicar la conducta sin prestar atención a la mente. Frente a cualquier tipo de

eliminativismo, nuestra postura es la del emergentismo, teoría según la cual la mente emerge del cerebro pero no actúa sobre él, sino que es un mero epifenómeno. Aún a pesar de las dificultades que plantea el estudio de la mente a causa de la mencionada exigencia de sintetizar los enfoques explicativo y comprensivo por la constitución biosocial de la mente, a nuestro juicio es una tarea imprescindible, en tanto que el reduccionismo materialista, como demostraremos mediante unos argumentos de Hilary Putnam, incurre en el error de presuponer la transitividad de las explicaciones.

Habiendo descrito las características generales de los paradigmas de la mente y del cerebro en los que se basan la IA simbólica y la IA subsimbólica, en el *capítulo quinto* aumentaremos el nivel de concreción. Lo haremos exponiendo un modelo de la inteligencia a nivel cerebral elaborado por un ingeniero informático, Jeff Hawkins, y un modelo cognitivista de la inteligencia a nivel mental elaborado por Roger Schank, un informático investigador de la IA que en los últimos tiempos ha pasado a dedicarse a la psicología. El modelo cognitivista de la inteligencia de Schank nos servirá para apreciar desde dentro las dificultades insuperables con las que se topa la IA simbólica en su intento por utilizar sistemas formales, que es lo que en el fondo son todas las computadoras electrónicas, para replicar la versatilidad distintiva de la inteligencia humana. Si el psicologismo epistemológico-lógico, como por ejemplo el de Stuart Mill, contra el que luchaba Husserl pretendía reducir la lógica a psicología, el cognitivismo pretende justo lo contrario: reducir la psicología a sistemas de lógica formal.

Hacia el final de la exposición del modelo de la inteligencia de Schank trazaremos un hilo conector con el tema del mito del método científico tratado en el capítulo previo. Nuestra tesis es que lo que subyace a la IA simbólica y al mito del método en el sentido antes precisado es la misma pretensión positivista: encontrar algoritmos generadores de teorías, ya sean científicas en el caso del método científico o precientíficas en el caso de la IA simbólica. En ambos casos se pretende automatizar la producción de conocimiento, una empresa imposible que, sin embargo, es signo de nuestro tiempo, y que está en continuidad con los propósitos para los que se crearon las redes de calculadores humanos y, más tarde, las computadoras electrónicas.

En el capítulo quinto expondremos también la teoría de las inteligencias múltiples (teoría IM) de Howard Gardner, la cual, frente a las teorías de la inteligencia de Hawkins y Schank, es la más acertada a nuestro juicio por dos razones principales. La primera es su caracterización relativista, en el mejor sentido, de la inteligencia como una facultad múltiple, distribuida y contextualizada, a diferencia de los enfoques unitarios, solipsistas y etnocéntricos que predominan en la psicometría, y de los cuales también hablaremos sobre la marcha. La otra razón de nuestro adhesión a la teoría de las inteligencias múltiples de Gardner es que los criterios que propone para identificar una inteligencia abarcan factores muy diversos, desde biológicos y culturales hasta históricos. Semejante amplitud refleja el carácter complejo de la inteligencia, en lugar de eludirlo como hacen aquellos psicólogos que definen la inteligencia en términos operacionales como aquella facultad que se mide con los tests de inteligencia.

Por supuesto, en una reflexión filosófica sobre IA y la naturaleza de la inteligencia no podía faltar un examen del test de Turing. A él dedicaremos la última sección del capítulo quinto. Lo expondremos, expondremos la refutación formulada contra él por John Searle con su famoso argumento de la sala china, expondremos las objeciones que se le han planteado al argumento de Searle y que han sido recopiladas por él mismo, expondremos las respuestas de Searle a dichas objeciones, y finalmente expondremos nuestras propias objeciones a esas respuestas. Nos sumergiremos, por tanto, en un diálogo de varias capas que culminará con una defensa por nuestra parte del criterio conductista de la inteligencia en el que se basa el test de Turing. Además evaluaremos la importancia que el matemático inglés otorgaba al lenguaje. Turing coincidía con Descartes, y nosotros con ellos dos, en que el lenguaje es una condición necesaria, y en cierto sentido también suficiente, de la inteligencia humana.

En el *capítulo sexto* repasaremos la Historia de la IA desde su fundación oficial en la conferencia de Dartmouth, celebrada en 1956, hasta el presente. En sus inicios la IA se adscribió al paradigma cognitivista, fundado oficialmente en un simposio del MIT que tuvo lugar en ese mismo año un mes después, y en el que participaron también dos de los asistentes que estuvieron en Dartmouth: Allen Newell y Herbert Simon, los

autores de la primera IA. O, más correctamente, el primer intento de IA, ya que el Logic Theorist, que así se llamaba el programa informático, no era una verdadera IA. De hecho, como veremos a lo largo del capítulo, en toda la Historia de esta disciplina jamás se ha logrado crear una IA en el sentido fuerte que coincide con la noción vulgar de IA. Newell y Simon, junto con otros destacados investigadores como Marvin Minsky, se aferraron desde el principio al programa de investigación de la IA simbólica, que es el de la duplicación de la mente entendida ésta en términos cognitivistas, al tiempo que empleaban su poder para hundir académicamente a aquellos que, como Frank Rosenblatt, se atrevían a incursionar en el enfoque de la IA subsimbólica, que es, recordemos, el de la duplicación de las redes de neuronas o, por lo menos, la simulación de su funcionamiento. La situación cambió hacia los años 80, y desde entonces ambos puntos de vista, simbólico y subsimbólico, coexisten e incluso se complementan. No obstante, ninguno de ellos ha conseguido construir una máquina tan inteligente como un ser humano. Cada programa de investigación presenta sus propios problemas, y en este capítulo los examinaremos. El principal de la IA simbólica es el de la limitación de dominio, es decir, la falta de versatilidad, que es, como venimos señalando, una de las dos características distintivas de nuestro intelecto. La otra, el lenguaje natural, tampoco ha podido ser replicada.

Aprovechando el hilo conductor de la Historia de la IA, expondremos también algunas de las estrategias, técnicas y arquitecturas más relevantes inventadas por los investigadores en su intento de crear máquinas pensantes. Esta exposición dará lugar a la presentación de alternativas opuestas tales como la consabida de IA simbólica vs subsimbólica, IA humana vs ajena, IA fuerte vs débil, IA abstracta vs situada, y métodos fuertes vs débiles. Se trata de conceptos que, aunque han sido acuñados con esos nombres por ingenieros y matemáticos, nos remiten a problemas filosóficos. Por ejemplo, la segunda IA de Newell y Simon fue el GPS, acrónimo de Resolutor General de Problemas, un programa diseñado, como su nombre indica, para resolver cualquier tipo de problema. Para lograr semejante proeza sus creadores optaron por un método débil, es decir, un método heurístico, de búsqueda de soluciones, que no dependía de

ningún conocimiento previo, frente a los métodos fuertes, que son aquellos que sí dependen de una base de datos que proporciona información útil para resolver los problemas. El método de Descartes para dirigir bien la razón y hallar la verdad en las ciencias era un método débil, pues el filósofo francés renunciaba por principio a confiar en la memoria, que es el equivalente humano a las bases de datos de las computadoras electrónicas. La Historia de la IA, como se aprecia este ejemplo, ha recorrido en su poco más de medio siglo muchas de las teorías del conocimiento que han sido propuestas por los filósofos a lo largo de dos mil quinientos años.

Dentro de la IA ha habido investigadores que han ido más allá de la técnica y han aprovechado la filosofía para identificar las causas profundas, epistemológicas, de ciertos problemas. A este respecto veremos lo que John McCarthy y Patrick Hayes señalaron en su artículo de 1969 *Some philosophical problems from the standpoint of artificial intelligence*. De entre todos los problemas apuntados en el texto, dos son los más relevantes: el problema del marco y el problema de la cualificación. El problema del marco refiere al problema de actualizar una base de datos interconectados entre sí, de tal manera que la modificación de tan sólo uno de ellos puede poner en marcha un efecto mariposa que resulte en la necesidad de modificar muchos otros. El problema de la cualificación, en cambio, es menos técnico y más filosófico. Por ser tal su naturaleza, será el problema irresoluble que emplearemos con preferencia en el capítulo siguiente, el séptimo, para refutar la posibilidad técnica de la IA simbólica. El problema de la cualificación refiere a la imposibilidad de elaborar un listado explícito de las condiciones de validez de las normas que los seres humanos empleamos para habérnoslas con el mundo, tanto en su dimensión física como social. Así, por ejemplo, operamos con la norma "los barcos sirven para navegar", pero ésta sólo es válida si se cumplen un montón de condiciones imposibles de explicitar, tales como que el pecio no debe tener un agujero en el casco. A su vez, el tamaño del agujero es una condición sujeta a condiciones tales como su tamaño, pues un agujero pequeño podría no ser un impedimento para la navegabilidad. De esta manera, las reglas tienes condiciones de validez que a su vez pueden remitir a otras en una sucesión indefinida.

Con toda la tarea de investigación realizada en los seis capítulos anteriores, en el *capítulo séptimo* estaremos ya en condiciones de abordar una de las dos grandes cuestiones del presente estudio: las condiciones de posibilidad técnicas de la IA. Al ser nuestro campo la filosofía, examinaremos dichas condiciones de posibilidad técnicas no desde el punto de vista de la ingeniería, sino de la epistemología. Adelantamos ya que el examen más interesante será el de las condiciones de posibilidad de la IA simbólica. No obstante, en el de la IA subsimbólica también trataremos asuntos de relevancia filosófica, tales como el de la relación entre la mente y el cuerpo. Para más detalles sobre el análisis epistemológico de la IA simbólica que efectuaremos en el capítulo séptimo nos remitimos a la sección dedicada a la metodología.

En cuanto al *capítulo octavo*, el último, su tema será el de las condiciones de posibilidad sociales de la IA. Para dirimir si esta tecnología tiene cabida en la sociedad actual, y para pronosticar con prudencia los usos más plausibles que se le darán, hemos de comenzar por un análisis de la propia sociedad. Lo haremos desde el punto de vista de la escuela de Frankfurt, y más concretamente, de Horkheimer, Adorno y Marcuse, acudiendo a las descripciones que ellos hicieron de la sociedad industrial avanzada y de su cultura de masas. Sin embargo, por tratarse de un tema metodológico, nos remitimos también a la sección de metodología. Para darle una cierta pátina distintiva de originalidad, y por continuar con el protagonismo que le hemos dado a la psicología en anteriores capítulos, en éste utilizaremos la teoría de la psicopatía del psicólogo Robert Hare. Qué tiene que ver esto con la sociedad actual y con la IA son preguntas que preferimos no contestar todavía.

Metodología

Para analizar las condiciones de posibilidad técnicas de la IA emplearemos un doble enfoque metodológico, pluralidad que viene impuesta por la diferencia ontológica entre la IA simbólica, que, como hemos señalado en la introducción, aspira a reproducir la mente tal y como ésta es concebida por el paradigma cognitivista de la psicología, y la IA subsimbólica, que tiene por objetivo la duplicación del cerebro. En cuanto a la IA simbólica, expondremos y ampliaremos la crítica realizada por Hubert Dreyfus, un filósofo que emplea una amplia batería de argumentos contra el modelo cognitivista de la mente. Dreyfus, contando con la colaboración puntual de su hermano Stuart, se hizo famoso en los años 60, justo cuando la IA acababa de constituirse como disciplina, por realizar una serie de ataques que le valieron la animadversión de todos los investigadores de este campo.

Los argumentos de Dreyfus deben su fuerza a que se basan en una concepción fenomenológica de la mente, es decir, molar, frente a la molecular sostenida por la IA y el cognitivismo. Dreyfus toma su enfoque molar de filósofos existencialistas como Heidegger y Merleau-Ponty y de psicólogos de la Gestalt como Max Wertheimer y Kurt Goldstein. Con estas herramientas, realiza una doble labor de destrucción de la IA simbólica y del cognitivismo. La primera consiste en señalar cuatro procesos cognitivos no explicables desde un enfoque molecular de la mente. Éstos son la periferia de la conciencia, la tolerancia a la ambigüedad, la discriminación de lo esencial y lo inesencial y la agrupación perspicaz. La periferia de la conciencia, más que un proceso, es el lugar situado más allá de la conciencia en el que tienen lugar procesos cognitivos de los que no tenemos noticia y sobre los que no tenemos control. Es allí donde el maestro de ajedrez ve la jugada clave sobre el tablero sin necesidad de calcular cientos

de secuencias de movimientos como hacen las computadoras, una habilidad que es posible gracias a la noción gestáltica de fondo y figura, pues el historial de partidas jugadas actúa como fondo que determina qué es lo que se presenta en la conciencia del maestro como una figura que atrae su atención. Los otros tres procesos cognitivos no replicables por la IA simbólica acontecen todos en la periferia de la conciencia. De este modo, tampoco tenemos conciencia de la manera en que toleramos la ambigüedad, discriminamos entre características esenciales a inesenciales y aplicamos la agrupación perspicaz, que es la habilidad en la que se basa el reconocimiento de patrones. Con esta crítica Dreyfus demuestra la imposibilidad por principio de la IA simbólica, pero sólo de la vertiente de la IA humana, que es la que tiene por objeto la replicación de la mente humana de una manera realista.

En cambio, seguiría en pie la posibilidad de la IA ajena, que se define como el enfoque que aspira a construir una mente cuya conducta resulte indistinguible de la humana aunque lo haga mediante procesos cognitivos diferentes. Para demostrar también la imposibilidad de la IA ajena, Dreyfus realiza una segunda labor de destrucción de la IA simbólica y del cognitivismo: indagar en los cuatro supuestos subyacentes a la IA simbólica en general. Éstos son, de menor a mayor importancia, el biológico, el psicológico, el epistemológico y el ontológico. El biológico sostiene que el cerebro es una máquina de estado discreto semejante a una computadora electrónica. El psicológico, también conocido como la hipótesis fuerte del sistema de símbolos (HFSS), postula que la mente humana utiliza procesos computacionales para producir la conducta inteligente, que es justo lo mismo que sostiene uno de los dos supuestos nucleares del cognitivismo: la metáfora computacional. El epistemológico, también conocido como la hipótesis del sistema de símbolos (HSS), no se pronuncia sobre cómo opera la mente humana, sino que, adquiriendo un menor grado de compromiso, tan sólo afirma que un sistema formal es suficiente para producir conducta inteligente en tanto que toda conducta inteligente puede ser formalizada. Y, por último, el supuesto ontológico es el que, en términos de la teoría de la ciencia de Lakatos, diríamos que está en el núcleo mismo del programa de investigación de la IA subsimbólica, y que,

por tanto, refutándolo a él, cae refutada toda la IA simbólica, tanto la humana como la ajena. El supuesto ontológico no es más que una forma de denominar al atomismo lógico de Bertrand Russell y del primer Wittgenstein: el supuesto de que el mundo entero puede ser representado como un conjunto estructurado de descripciones reducibles a expresiones primitivas o atómicas.

La labor de demolición de Dreyfus es progresiva. Comienza desmontando el supuesto biológico y termina por refutar el supuesto ontológico. Por nuestra parte, ampliaremos la crítica de Dreyfus de dos formas. La primera, acumulando evidencias tomadas de otras fuentes, para respaldar con mayor firmeza si cabe sus argumentos. Esto lo haremos refiriéndonos a asuntos que fueron comentados en anteriores capítulos, y no por casualidad, sino precisamente con el objetivo de hacerlos converger aquí. Así, por ejemplo, apelaremos a la exposición de los elementos de neurociencia del capítulo cuarto para reforzar la refutación del supuesto biológico, o rescataremos el tema de la estructura circular de la comprensión según Heidegger y Gadamer visto en el capítulo quinto para argumentar contra el objetivismo del atomismo lógico.

La segunda forma en que ampliaremos la crítica de Dreyfus será con una sección en la que enfrentaremos los argumentos pragmatistas del segundo Wittgenstein contra los argumentos atomistas del Wittgenstein del *Tractatus*. Dado que la IA simbólica se sustenta en la ontología del atomismo lógico, quién mejor que el propio Wittgenstein, el mayor exponente de esa doctrina, para acabar con ella. Rescataremos los momentos más brillantes de las *Investigaciones filosóficas* y de los *Cuadernos azul y marrón* en lo que el filósofo alemán arremete contra sus creencias anteriores. A este respecto el modelo de la mente de Roger Schank, expuesto en el capítulo quinto y retomado aquí, resultará esclarecedor, ya que se trata de un modelo cognitivista, basado por tanto en la manipulación regulada de expresiones simbólicas, pero al mismo tiempo fuertemente influido por el pragmatismo, como se aprecia en sus frecuentes alusiones a John Dewey y en su intento de explicar las habilidades prácticas de la vida ordinaria. Por no estar formado en la filosofía, Schank no se da cuenta de que está intentando acomodar dos ontologías incompatibles: la del

atomismo lógico que subyace al cognitivismo por un lado y la del pragmatismo por el otro. El cognitivismo se fundamenta en el supuesto de que la mente es como una computadora electrónica, y como tal ha de operar ejecutando reglas inequívocas en todo momento. Sin embargo, como observa el segundo Wittgenstein, la mente no siempre opera calculando sobre reglas, sino que, simplemente, hay cosas que son como son, y que sabemos hacer sin poder explicitarlas de manera algorítmica, y aun cuando es posible explicitarlas, tal formulación no es suficiente para ejecutarlas. El problema de la cualificación de McCarthy y Hayes, al que nos hemos referido en la introducción, es de una relevancia máxima en este punto, dado que señala que las reglas descansan sobre condiciones de validez imposibles de explicitar, una explicitación que, sin embargo, es imprescindible para las computadoras electrónicas.

En cuanto a la IA subsimbólica, compartimos con Dreyfus la valoración de que es técnicamente posible siempre y cuando adopte el enfoque de la IA situada, que es, frente al de la IA abstracta, el que pretende crear máquinas pensantes situándolas en un cuerpo. Sin embargo, en vez de argumentar desde la filosofía del cuerpo de Merleau-Ponty como hace Dreyfus, nosotros lo haremos desde la neurociencia. Como dice el neurocientífico Antonio Damasio, creer en la posibilidad de replicar el cerebro humano sin el cuerpo sería incurrir en una especie de cartesianismo materialista. Para estimar la cercanía en el tiempo de la construcción de una IA subsimbólica examinaremos las más modernas técnicas de modelado de redes de neuronas, tales como la evolución artificial. El mayor desafío de la IA subsimbólica es descifrar el conectoma de partida del cerebro humano, es decir, el cableado inicial de las redes de neuronas determinado por factores genéticos y sobre el cual la experiencia imprime las modificaciones que resultan en el cerebro adulto inteligente.

Respecto a las condiciones de posibilidad sociales de la IA, nuestro enfoque será el de la escuela de Frankfurt. En concreto, para desvelar los intereses de la sociedad actual hacia la IA utilizaremos las teorías de Horkheimer, Adorno y Marcuse sobre la sociedad industrial avanzada y la cultura de masas. La Ilustración, entendida como el proceso histórico de racionalización para liberar al hombre del miedo y erigirlo

en señor, ha resultado paradójicamente, como señalan Horkheimer y Adorno, en una recaída en el terror y la esclavitud. La razón subjetiva o instrumental, que es la creadora de la ciencia moderna y de la técnica, se ha liberado de la razón objetiva o axiológica, que es la rectora de las acciones del hombre, por no poder ésta demostrar una eficacia pragmática tan elevada como aquélla. Nuestra aportación a esta filosofía de la historia consistirá, principalmente, en esclarecerla trazando el perfil psicológico del individuo resultante de ella. Dicho perfil se caracteriza, a nuestro juicio, por su semejanza con dos patologías de la mente: la psicopatía y la esquizofrenia.

El psicópata es el triunfador, el hombre de éxito de nuestro tiempo, un tiempo que, al medirse en términos puramente instrumentales, sólo conoce el éxito como criterio de racionalidad. Para describir los rasgos de la psicopatía utilizaremos la teoría del psicólogo Robert Hare, considerado como el mayor experto mundial en la materia. Así como Horkheimer y Adorno emplearon en su *Dialéctica de la Ilustración* a los personajes del Marqués de Sade para retratar al individuo guiado por la sola razón subjetiva, nosotros emplearemos la descripción del psicópata según el doctor Hare, obtenida mediante la psicología experimental y la psicofisiología cognitiva. Por lo menos desde un punto de vista filosófico, y quizás también del diagnóstico clínico, los cuatro libertinos de *Las ciento veinte jornadas de Sodoma* tienen el perfil de un psicópata: son egocéntricos, sin rastro alguno de sentimiento de culpa, carentes de empatía, manipuladores, fríos, impulsivos, adictos a las experiencias extremas.

Sin embargo, no se muestran así todo el tiempo, porque saben que, si lo hicieran, no podrían llevar a cabo sus planes con éxito. En la práctica, tales conductas son las que funcionan en la sociedad, pero formalmente son rechazadas por una serie de valores completamente opuestos que, en virtud de tal oposición, generan esquizofrenia. Mientras se hace una cosa, se dice la contraria. En la obra de Sade *Los infortunios de la virtud*, la desdichada Justine era brutalmente explotada por personajes que en público sabían dar la apariencia de virtud que la sociedad exige. Una doctrina ésta, la de la doble verdad, que genera esquizofrenia, y que es imprescindible para alcanzar el éxito en la sociedad actual.

La IA, en tanto que tecnología automatizadora de la producción y por tanto reductora del tiempo de trabajo socialmente necesario e individualmente represivo, contiene una potencia revolucionaria para trascender semejante estado de cosas. Sin embargo, la actualización de tal potencia no acontecerá de manera necesaria como sostiene el positivismo en virtud de unas supuestas leyes del progreso social, sino que, desde un punto de vista marxista y antimecanicista como es el de la escuela de Frankfurt, depende de la acción de un sujeto histórico que debe estar concienciado de que tal acontecer es posible desde el punto de vista de la razón subjetiva y necesario desde el de la razón objetiva. La toma de esa conciencia es, no obstante, imposible a causa de la acción totalitaria a escala mundial de la cultura de masas, tal y como demuestran Horkheimer y Adorno. El cierre del universo del discurso y de la acción operado por la industria cultural es tan perfecto que ni siquiera puede ser negado dialécticamente mediante la astucia de Odiseo, la esencia de la razón subjetiva cuyo abuso ha hecho fracasar a la Ilustración y nos ha arrojado a la situación presente. En consecuencia, mientras la industria cultural funcione con la eficacia exhibida hasta ahora, el avance de la Historia permanecerá suspendido aun a pesar de que surjan tecnologías con un potencial revolucionario tan alto como la IA.

Para terminar esta sección dedicada a la metodología hemos de hacer tres aclaraciones más. La primera es que, a menos que se indique lo contrario, utilizaremos la expresión "psicología cognitiva" para designar no a aquella parte de la psicología que se ocupa del estudio de la cognición, sino para referirnos al paradigma cognitivista de la psicología. La segunda es que utilizaremos el anglicismo "cognitivo", en lugar de la palabra correcta en nuestro idioma, que es "cognoscitivo". Tal decisión obedece a razones de claridad expositiva, ya que en la bibliografía empleada los traductores al español suelen preferir el anglicismo. Y la tercera aclaración tiene que ver justamente con la bibliografía, y es que para la citación utilizaremos las normas más recientes de la guía de estilo y formatos de la APA, la American Psychological Association, pero con algunas modificaciones, siendo la principal que, en lugar de emplear como índice la fecha de la edición manejada, emplearemos la de la primera edición de la edición

manejada. Así, por ejemplo, la obra de Kuhn *La estructura de las revoluciones científicas* apareció por vez primera en 1962, pero en la edición de 1970 se añadió una posdata importante, razón por la cual utilizaremos la fecha 1970. La razón de esta decisión es proporcionar al lector un índice, el de la fecha de la obra, que sea significativo, en tanto que permite ubicarla temporalmente, mientras que la fecha de la edición manejada no aportada nada al contenido de la lectura. En cuanto a los textos clásicos, entendiendo por tales arbitrariamente los anteriores al siglo XX, en lugar de citarlos mediante fecha lo haremos por su nombre.

1. La IA en el imaginario popular

Las inteligencias artificiales (IA) pueden ser de dos tipos: fuertes y débiles (Searle, 1980, p. 282). Las *débiles* son modelos informáticos de ciertos procesos mentales (IA simbólica) o cerebrales (IA subsimbólica) que se diseñan con el propósito de que resulten útiles para el estudio científico. Las *fuertes* son también modelos informáticos, pero que simulan la mente o el cerebro en su totalidad (IA humana) o bien sólo la conducta producida por ellos (IA ajena). La IA fuerte, y no la débil, es justamente la *noción vulgar de IA* que se maneja a pie de calle. Pregúntesele a cualquiera lo que es una IA y responderá con una definición aproximada a ésta de IA fuerte: una máquina con una inteligencia parecida a la de un ser humano. Por su condición de máquina se espera que sirva a su dueño, y por su condición de inteligente se espera que lo haga con unas destrezas intelectuales parecidas a las de un ser humano. Ciertamente, si una máquina se comportase de manera indistinguible de como lo haría una rata, también sería calificada de inteligente, pues la inteligencia no es un atributo booleano de todo o nada, sino gradual (Franklin, 1995, p. 17), y las ratas poseen un cierto grado de inteligencia. Sin embargo, en el presente estudio sólo vamos a explorar las condiciones de posibilidad técnicas y sociales de la recreación artificial de la inteligencia humana. Así pues, cuando hablemos de inteligencia sin más nos estaremos refiriendo siempre a la del ser humano, y cuando hablemos de IA nos estaremos refiriendo sólo a la recreación artificial de la inteligencia humana.

El hecho es que, en la actualidad, las inteligencias artificiales fuertes no existen, y no está claro si existirán algún día. Por tanto, su presencia en el imaginario popular no puede explicarse por el contacto con ellas, sino que es producto de un proceso de construcción social llevado a cabo por tres agentes principales: la mercadotecnia, los

propios investigadores de la IA y la ciencia ficción. En este primer capítulo vamos a describir la noción vulgar de IA y a desvelar los intereses por los cuales dichos agentes la han construido tal y como es, es decir, identificándola con la IA fuerte.

Definición preliminar de la inteligencia

Para entender lo que es una IA en sentido vulgar es necesario fijar primero qué es la inteligencia humana, y a continuación determinar en qué consiste el parecido que debe tener la artificial respecto de la humana. Por desgracia, la psicología contemporánea no ofrece una definición universalmente aceptada de la inteligencia (Davidson & Kemp, 2011, p. 58). De hecho, en el terreno de la inteligencia, y de la psicología en general, casi todo son arenas movedizas. La ausencia de certeza, o cuando menos de consenso, es habitual. A falta todavía de una definición más extensa como la que desarrollaremos en el capítulo quinto, una preliminar bastante buena es la de Donald Michie, que captura con audacia los rasgos compartidos por las diversas definiciones de inteligencia que circulan por la literatura especializada. Según Michie: «El intelecto humano se señala, no tanto por su especial brillantez en una tarea particular, como por su capacidad de hacer un intento pasable en casi cualquier cosa» (Michie, 1974, p. 51). Lo esencial de la inteligencia humana es, por tanto, la capacidad de habérselas pasablemente con casi cualquier problema.

En la década de los 50 unos cuantos pioneros de la informática como Donald Michie pusieron en marcha el proyecto de construir inteligencias artificiales. Desde entonces disponemos de máquinas, típicamente computadoras electrónicas, que realizan de forma automática tareas que, si fueran realizadas por un ser humano, requerirían inteligencia. Sin embargo esto no quiere decir que sean inteligentes, porque no es lo mismo ser inteligente que ser capaz de realizar una tarea que requeriría inteligencia si la realizara un ser humano. Una calculadora de bolsillo, por ejemplo, realiza operaciones aritméticas que requerirían inteligencia si las realizara un ser humano, pero sería una extravagancia considerar que una calculadora es una

inteligencia artificial. Se puede defender que una calculadora es inteligente (Franklin, 1995, p. 107), del mismo modo que también se puede defender que un termostato tiene conciencia de la temperatura ambiente. Pero en ambos casos se trata de opiniones marginales, excéntricas, que no se ajustan a la noción vulgar de IA.

Partiendo de la definición de la inteligencia según Michie, podemos abordar la cuestión del parecido que debe guardar una inteligencia artificial con una humana. Para que una máquina merezca el calificativo de inteligente ha de ser capaz de realizar no una, ni dos, ni un número definido de tareas que requerirían inteligencia si las realizara un ser humano, sino que debe ser capaz de realizar casi cualquier tarea que requeriría inteligencia si la realizara un ser humano. Naturalmente, si pudiera realizar no casi, sino cualquier tarea de ese tipo e incluso alguna más como por ejemplo resolver raíces cuadradas de diez dígitos en menos de un segundo, entonces también sería una IA. Pero eso sería una condición deseable. La *condición necesaria y suficiente* de una IA en sentido vulgar es que sea capaz de *casi* cualquier cosa propia de la inteligencia humana. Y la inteligencia humana se define, a su vez, por su capacidad de hacer un intento pasable en *casi* cualquier cosa. Por tanto, para que una máquina sea una IA debe ser capaz, como mínimo, del "casi del casi".

Deficientes geniales

Una calculadora no está ni siquiera cerca de ser casi tan inteligente como un ser humano, sencillamente porque no es capaz de hacer un intento pasable en casi cualquier cosa. Por tanto, no es una IA. No obstante, en el ámbito de las operaciones aritméticas hay que reconocerle una competencia superior a la del ser humano promedio, pues poca gente hay capaz de resolver mentalmente raíces cuadradas de diez cifras, y desde luego no hay nadie que lo haga en una fracción de segundo y sin margen de error, esto es, con tanta velocidad y precisión. Como la calculadora, hay muchas otras máquinas que realizan tareas intelectuales mejor de como lo haría un ser humano, incluso tareas que no podrían ser realizadas ni por toda la humanidad entera.

Dentro de este tipo de máquinas, las más similares a la inteligencia humana en la actualidad son los *sistemas expertos* (*expert systems*) (Jackson, 1986, p. 1). La diferencia entre un sistema experto y una verdadera IA fuerte es que un sistema experto es excepcionalmente competente sólo en problemas dentro de un ámbito restringido, mientras que una IA fuerte debe tener la excepcional capacidad de abordar problemas de casi cualquier ámbito, pues el rasgo esencial de la inteligencia es, como señala Michie, su capacidad para habérselas con casi cualquier cosa.

Uno de los sistemas expertos más célebres de todos los tiempos es Deep Blue, el superordenador construido por IBM que en 1997 derrotó al entonces campeón del mundo de ajedrez Gary Kaspárov. Para los estándares de aquella época, Deep Blue era un prodigio de la tecnología, capaz en un solo segundo de efectuar 11.000 millones de operaciones y calcular 200 millones de posiciones. La máquina ganó a Kaspárov en un encuentro al mejor de seis partidas. Sin embargo, al término de la última partida, el único que compareció ante la prensa fue Kaspárov, porque Deep Blue, sencillamente, no tenía ninguna capacidad de hablar, ni con los periodistas ni con nadie. Deep Blue sabía jugar al ajedrez, y demostró que lo hacía mejor que el mejor de los humanos, al igual que una modesta calculadora de bolsillo calcula mejor que el mejor de los humanos. Pero Deep Blue no sabía hacer nada más, ni siquiera charlar sobre su único dominio de conocimiento, el ajedrez.

Los sistemas expertos son en el reino de las máquinas el equivalente a los *deficientes geniales* o *síndrome de savant* en la especie humana: sujetos que realizan unas pocas tareas de forma genial pero que en el resto son deficientes. En cambio, sería tolerable que una IA fuese mediocre jugando al ajedrez, siempre y cuando fuera capaz de acometer casi cualquier tarea intelectual con un nivel de destreza parecido al de un ser humano normal. Hasta la fecha ninguna máquina ha superado el abismo que separa a los sistemas expertos de las inteligencias artificiales, un abismo que es incluso mayor que el que separa a los deficientes geniales de las personas normales, pues los deficientes geniales suelen saber hablar. He aquí la razón por la cual ninguna máquina de las que existen en la actualidad merece, en rigor, ser calificada de inteligente.

Mundo físico y mundo social

Pretender alcanzar la polivalencia de la inteligencia sumando sistemas expertos sería como pretender alcanzar la Luna añadiendo peldaños a una escalera. El número de problemas que un ser humano puede resolver a lo largo de su vida es potencialmente infinito, y al infinito es imposible llegar sumando uno más uno. Aunque se lograra la titánica empresa de integrar en una sola máquina todos los sistemas expertos más avanzados que existen, esa máquina seguiría siendo el equivalente a un deficiente genial en grado extremo, pues tendría gravísimos problemas de comunicación con el mundo en sus dos dimensiones: la física y la social. Para habérselas con el *mundo físico* de manera eficaz es condición necesaria el reconocimiento de objetos, mientras que en el *mundo social* el requisito imprescindible es comprender el lenguaje natural. Sin embargo, ninguna máquina hasta la fecha ha demostrado la competencia suficiente en ninguna de las dos tareas. Hasta una rata común, con su diminuto cerebro de dos gramos, es capaz de reconocer objetos de forma *rápida* y *precisa*, mientras que las computadoras electrónicas necesitan procesar millones de ciclos para alcanzar resultados llenos de errores. Las computadoras electrónicas son rápidas y precisas en algunas tareas como resolver raíces cuadradas, pero *lentas* y *torpes* en otras.

En cuanto a la comprensión del lenguaje natural, el fracaso de la IA también es palmario. En 1966, Joseph Weizenbaum, profesor de ingeniería informática del MIT, publicó un programa de ordenador llamado ELIZA (Weizenbaum, 1976, p. 14). Bautizada en honor al personaje homónimo de la obra de teatro *Pigmalión* de Bernard Shaw, ELIZA era capaz de conversar con un ser humano mediante un teclado y un monitor. En cuestión de meses, sobre la base de ELIZA, el psiquiatra Kenneth Colby desarrolló el programa DOCTOR, cuya utilidad era realizar psicoterapia rogeriana, la cual se caracteriza por que el terapeuta no cesa de hacer preguntas al paciente para que éste llegue así, casi por sí solo, a descubrir el origen de sus problemas. DOCTOR

causó un gran impacto en el ámbito académico de la época, y enseguida empezó a circular en forma de copias por varias universidades de los Estados Unidos. Desde los becarios hasta los jefes de departamento, todo el mundo deseaba sentarse frente a alguna de las escasas computadoras del campus para contarle a DOCTOR sus secretos más íntimos. Varios psiquiatras, encabezados naturalmente por el propio Colby, llegaron incluso a proclamar públicamente que DOCTOR debía ser comercializado. Su mayor atractivo, argumentaban, era que un solo ordenador con una copia de DOCTOR instalada en el disco duro podría proporcionar psicoterapia a cientos de usuarios simultáneamente. El reducido coste del dispositivo permitiría que las personas sin recursos suficientes para pagar una psicoterapia convencional pudieran acceder a ella de esta manera alternativa.

Lejos de sentirse feliz por las alabanzas hacia su creación, Weizenbaum reaccionó encolerizado. Por un lado, expuso argumentos morales en virtud de los cuales una máquina no debía sustituir jamás al ser humano en ciertas tareas. Y por otro, recalcó argumentos técnicos que demostraban que ELIZA y DOCTOR eran incapaces de comprender verdaderamente a su interlocutor humano. Ambos programas conseguían mantener conversaciones gracias a un ingenioso sistema de *scripts*, o reconocimiento de patrones, similar al guión que utiliza un actor para dar la réplica aunque el texto esté escrito en un idioma que no entienda. No obstante, la argucia empleada por Weizenbaum no era suficiente para superar las numerosas dificultades epistemológicas que entraña la reproducción artificial de la comprensión y, en consecuencia, con un poco de paciencia frente al teclado, el usuario terminaba descubriendo que el supuesto terapeuta que había al otro lado del monitor no era más que una máquina incapaz de comprender lo que se le decía. Cierto es que ELIZA y otros programas basados en técnicas similares, como SAM de Roger Schank y SHRDLU de Terry Winograd, datan de los inicios de la IA, pero en la actualidad la comprensión del lenguaje natural continúa siendo un problema para las máquinas. Prueba de ello es la irritación que experimentamos al intentar hablar por teléfono con un contestador automático "inteligente".

En definitiva, las inteligencias artificiales en sentido vulgar no existen. Sin embargo, se trata de un concepto usado a pie de calle. Por tanto, debemos preguntarnos de dónde procede, dado que no es por el contacto con algo real. La respuesta, a nuestro juicio, es que la noción vulgar de IA se trata de un constructo social que ha sido incorporado al imaginario popular en el último medio siglo por la acción de tres agentes principales: la mercadotecnia, los propios investigadores de la IA y la ciencia ficción. En las secciones que siguen vamos a describir la conducta de estos tres agentes y a desvelar los intereses que los han movido a participar en la construcción social de la noción vulgar de IA. Hay otra cuestión a este respecto que debe ser examinada, y es el motivo por el cual el público se siente tan atraído por las máquinas inteligentes. Pero ese tema lo dejamos para el capítulo segundo.

1.1. Mercadotecnia

La mercadotecnia utiliza el atributo "inteligente" de manera procelosa como gancho comercial. Una y otra vez, los carteles publicitarios aseguran que ya podemos comprar máquinas inteligentes: televisores y teléfonos, por ejemplo. Los llaman *smart TVs*, *smartphone*s: reparemos en que *smart* en inglés significa "inteligente". Se trata de una mentira repetida muchas veces y, en cierto sentido, una mentira repetida muchas veces termina convirtiéndose en una verdad (Nietzsche, *Sobre verdad y mentira en sentido extramoral*, p. 42). Pero una verdad de semejante tipo no es más que un constructo social, y como tal puede ser deconstruido para que la mentira reaparezca. El grado de verdad de una cosa (*res* en latín) es proporcional a su *res*-istencia contra los esfuerzos por superarla o deconstruirla (Woolgar, 1988, p. 91). Aplicando este principio se descubre que el grado de verdad de la IA en sentido vulgar es pequeño, pues deconstruirla es fácil. La mercadotecnia es el epítome de la construcción de la verdad mediante la repetición de la mentira. Con el objetivo de aumentar las ventas, el fabricante pretende convencer al consumidor de que un producto es inteligente, y para lograrlo no tiene más que repetírselo miles de veces en todas partes.

Las construcciones, tanto del mundo físico como del mundo social, se deconstruyen de dos maneras: intencionalmente a martillazos, como decía Nietzsche que él filosofaba (Nietzsche, *Ecce homo*, p. 313), o dejando que pase el tiempo. A través del tiempo, la ley de la entropía actúa como una fuerza ciega imparable que destruye todas las formas, desde las naturales como las estrellas, hasta las artificiales como las pirámides levantadas por los hombres. Mirando hacia la constelación de Orión, la Gran pirámide de Guiza lleva seis mil quinientos años en pie. En cambio, los montajes de la mercadotecnia no aguantan ni unas décadas, y a veces ni siquiera unos meses o semanas. Así, hoy en día nos parecen ridículos los anuncios de televisión en blanco y negro en los que un ama de casa declara estar encantada con su nuevo robot de cocina "inteligente". Del mismo modo, a la próxima generación le parecerán ridículos los reclamos publicitarios actuales en los que se califica de inteligente a un teléfono móvil sólo porque es capaz de realizar unos cuantos trucos como, por ejemplo, obedecer órdenes sencillas por reconocimiento de voz. Por mucho que se repita en los medios de comunicación que una determinada máquina es inteligente, la verdad es que ninguna lo es. Lo más parecido que hay son los sistemas expertos, es decir, los deficientes geniales del reino de las máquinas.

El éxito del término "inteligente" como gancho comercial es tan formidable que los fabricantes no se conforman con aplicarlo a las computadoras electrónicas o, por lo menos, a mecanismos complejos, sino que van más allá, y ya hablan de casi cualquier cosa como si estuviera dotada de inteligencia. Así, dicen que venden materiales de construcción inteligentes, cuando en realidad lo único que tienen de especial es que recuperan su forma de fábrica tras una deformación o cambian su densidad ante ciertas condiciones ambientales, o dicen que un tejido es inteligente sólo porque las fibras están agrupadas de manera tal que permiten la transpiración de la piel. Si creyéramos lo que proclaman los publicistas, resultaría que el mundo es un lugar en el que cada vez más abunda la inteligencia, a pesar de las guerras, la pobreza y la destrucción de los ecosistemas. Y en cierto sentido dicen la verdad: el mundo es cada vez más racional. Pero racional en sentido puramente instrumental.

La *razón instrumental* es aquella que se ocupa de maximizar el rendimiento de los medios para alcanzar un fin, sin cuestionar la racionalidad del fin en sí mismo. De esta manera, el sufrimiento de millones de seres humanos y la aniquilación a escala global de las reservas naturales son totalmente racionales en sentido instrumental, en tanto que son fenómenos planificados para la consecución optimizada de un fin, que es el enriquecimiento de las corporaciones empresariales. El mundo presente es de una racionalidad instrumental superior a la de cualquier otra época gracias a varios factores, entre los cuales las computadoras electrónicas ocupan un lugar destacado. De momento no son tan inteligentes como los vendedores anuncian, pero si algún día llegan a serlo, su aportación a la racionalización instrumental del mundo será aún mayor. Profundizaremos sobre este asunto en el capítulo octavo, dedicado a las condiciones de posibilidad sociales de la IA fuerte.

1.2. Investigadores de la IA

En 2012 la PNAS (*Proceedings of the National Academy of Sciences*) publicó un estudio demoledor que revelaba que en el campo de la biomedicina y las ciencias de la vida el número de artículos fraudulentos se había multiplicado por diez desde 1975 (Fang, Steen & Casadevall, 2012). El estudio, realizado conjuntamente por la Universidad de Washington y la Facultad de Medicina Albert Einstein de Nueva York, se basaba en el seguimiento de 2.047 artículos retractados. De ellos, sólo un 21,3% fueron atribuibles a errores, mientras que el 67,4% restante se debía a mala conducta. Dentro de los artículos producidos por mala conducta, el 43,4% había sido retirado por fraude o sospecha de fraude, el 14,2% eran duplicaciones, y el 9,8% eran directamente plagios. La mala conducta es una práctica que se extiende a todos los campos de la ciencia, y en el caso de la IA, es hasta proverbial (Dreyfus, 1992, p. 151).

No es sorprendente que los fabricantes de electrodomésticos mientan por dinero. Lo que sí es sorprendente es que los científicos mientan. O, al menos, es sorprendente para el hombre de la calle, que tiene una *concepción mertoniana de la*

ciencia, según la cual la actividad científica se diferencia de cualquier otra actividad humana porque está regulada por un conjunto de imperativos morales que garantizan la pureza racional del resultado (Merton, 1973, p. 269), entendiendo por racional lo excluyente de lo pasional. Esos imperativos se conocen habitualmente por su acrónimo *CUDOS*, que son las siglas en inglés de comunalismo, universalismo, desinterés y escepticismo organizado. El *comunalismo* (*communism*) refiere a que los productos de la inteligencia humana, en este caso las teorías científicas, suelen resultar de la colaboración entre varios sujetos, ya sea directamente en persona, o indirectamente aprovechando obras ajenas para el beneficio de las propias, que es lo que sucede cuando, por ejemplo, un astrónomo utiliza el cálculo integral, inventado por Newton, para realizar las operaciones que le servirán para demostrar una nueva teoría. El *universalismo* (*universalism*) significa que la objetividad excluye el particularismo, y en consecuencia la aceptación o el rechazo de una teoría científica no depende de los atributos personales o sociales de su autor, tales como el prestigio. El *desinterés* (*disinteredness*) se refiere a la obligación moral de anteponer el bien común de la ciencia en particular y de la humanidad en general a cualquier beneficio individual, dice Merton, obtenible a través de prácticas deshonestas como el fraude, las añagazas y las reivindicaciones irresponsables. Y, por último, el *escepticismo organizado* (*organized skepticism*) apunta al compromiso de mantener una actitud prudente, de no dejarse llevar por lo que se hace y lo que se dice, sino reflexionar con criterio propio antes de formarse una opinión.

La IA es una disciplina científica que, en su medio siglo de andadura, ha estado siempre situada en las antípodas del universalismo, del desinterés y del escepticismo organizado; sólo le ha faltado incumplir el imperativo del comunalismo, pero no ha podido hacerlo porque, como veremos en el capítulo quinto a propósito de la teoría de las inteligencias múltiples de Howard Gardner, se trata de una forma alternativa de referirse al carácter distribuido de la inteligencia, y éste es un rasgo inevitable de la inteligencia humana. El filósofo Jack Copeland denuncia que: «La cantidad de engaños públicos de la IA no tiene paralelo en los anales de los estudios académicos»

(Copeland, 1993, p. 148). Él lo denomina "inflar": exagerar, abultar hechos, noticias. Cada poco tiempo, los medios de comunicación bombardean al público con un supuesto gran avance en IA o con las declaraciones de un científico asegurando que estamos muy cerca del día en que las máquinas serán tan inteligentes como para hacer todo nuestro trabajo.

Ante esta información, la mente del hombre de la calle es una tabla de cera incapaz de oponer resistencia crítica. Su perfil intelectual es el de un sujeto que acostumbra a creer todo lo que dicen los rotativos de su inclinación ideológica. Y cuando se trata de cuestiones científicas, entonces cree incluso lo que dicen los medios de línea editorial opuesta, pues está ciegamente convencido de que los científicos nunca mienten, sino que se conducen de forma mertoniana, es decir, puramente racional sin dejarse perturbar por las pasiones. Además, a esta credulidad hay que sumarle el hecho de que los avances de la ciencia lo seducen con facilidad, pues en el fondo de su alma anida la ilusión positivista de que las aplicaciones prácticas de la ciencia devolverán a la humanidad a la felicidad originaria del Edén. El resultado de esta combinación de factores psíquicos es que las mentiras de los investigadores de la IA han calado hondo en el hombre de la calle, y ello ha contribuido a construir la noción vulgar de inteligencia artificial: siempre estamos a las puertas de una generación de máquinas totalmente inteligentes que no tienen ningún problema ni con el reconocimiento de objetos ni con el lenguaje natural.

Lucha por la supervivencia

Los científicos de la IA llevan décadas mintiendo, y lo hacen por el mismo motivo que las corporaciones empresariales: por dinero. Atrás quedó el período *amateur* de la ciencia en el que un solo hombre en una alcoba era capaz de realizar un descubrimiento importante. Eso fue así aproximadamente entre 1600 y 1800 (Woolgar, 1988, p. 30). Después hubo una fase *académica*, que duró hasta 1940, durante la cual la investigación científica se trasladó a las universidades por la

necesidad de contar con medios técnicos inasequibles para el investigador independiente. Y, por último, la ciencia entró en la fase actual, la *profesional*, en la cual la investigación se ha hecho tan costosa que las universidades necesitan atraer financiación privada del ámbito empresarial. Pero la financiación es limitada, y la limitación da lugar a que en el seno de las universidades y de otros centros de investigación se libre una lucha despiadada por los recursos.

En medio de la guerra de todos contra todos por el dinero, hasta el investigador mejor intencionado se ve obligado en ocasiones a prometer que el resultado de su proyecto será la piedra filosofal, aun a sabiendas de que no va a ser así. Y cuando el proyecto ha concluido con un resultado decepcionante, entonces su autor sigue defendiendo que es lo que en realidad no es, con el objetivo de que la mentira repetida muchas veces termine convirtiéndose en verdad, tal y como se hace en la mercadotecnia. Ante el dilema de mentir o fracasar en su carrera, los científicos recuerdan que son seres de carne y hueso con facturas en el buzón, y no suelen dudar en saltarse la ética mertoniana. Drew McDermott, de la Universidad de Yale, reconoce que, debido a la persistente actitud deshonesta de sus colegas, la IA ha estado siempre en el filo de la respetabilidad (McDermott, 1976, p. 143). En el mejor de los casos, dice McDermott, la mentira se debe a que el investigador está henchido de un inocente optimismo que le hace pensar que su proyecto supondrá un gran avance. Semejante conducta incumpliría el imperativo mertoniano del escepticismo.

Daniel Crevier, un empresario informático que ha estudiado a fondo la Historia de la IA, cuenta una anécdota que le fue relatada por Berthold Horn, profesor del MIT: «[Recuerdo el caso de] un viejo investigador que estaba a cargo de la conferencia de IA en Boston hace unos años. Había por allí muchos periodistas, y él andaba diciendo cosas como: "Dentro de cinco años tendremos robots que irán por nuestras casas recogiendo las cosas que hay tiradas por el suelo". Yo lo llevé a una esquina y le dije: "¡No hagas esas predicciones! Es algo que a la larga te meterá en problemas, como ya le ha sucedido a otras personas. Estás subestimando el tiempo que llevará conseguir esos robots". Él me respondió: "No me importa. Date cuenta de que las fechas que he

escogido son de cuando yo ya estaré retirado". Entonces yo le dije: "Vale, pero yo no estaré retirado, y la gente vendrá y me preguntará por qué no hay un robot en sus casas recogiéndoles los calcetines sucios"» (Crevier, 1993, p. 6). Es difícil decidir cuál de los dos comportamientos es más inmoral. El investigador viejo actúa mal porque miente sin rubor ante los medios de comunicación para obtener sus quince minutos de fama, pero es que al joven esto le trae sin cuidado, porque lo único que a él le preocupa es que las mentiras del otro lo metan en problemas a él. Parece una escena de *El Buscón*, que así llamaban a don Pablos, dice Quevedo, por ser un buscavidas capaz de cualquier bajeza para procurarse su propio bien.

Al margen del interés económico y del afán de notoriedad, Crevier señala otros tres factores importantes como causas de la conducta antimertoniana característica de la IA (Ibíd., p. 4). El primero es que el progreso de las computadoras electrónicas en las décadas de los 50, 60 y 70 fue tan rápido, que en aquel entonces era razonable pensar que, si la tasa de evolución exponencial en la informática se mantenía, las inteligencias artificiales en sentido fuerte tardarían poco en llegar. El segundo factor es el modelo simplista de la mente divulgado en aquella época por el emergente paradigma cognitivista de la psicología. Nos ocuparemos en profundidad del cognitivismo en el capítulo cuarto, así como de su influencia perniciosa sobre la IA en el capítulo séptimo, pero adelantamos ya que esta corriente en sus inicios consideraba que la mente era un procesador de información similar a una computador electrónica, dando lugar, por tanto, a una relación circular improductiva entre la IA y la psicología. Puesto que los ordenadores disponibles ya eran como cerebros electrónicos, sólo había que encontrar el programa adecuado que los hiciera funciona como inteligencias artificiales en sentido fuerte. El tercer y último factor adicional reseñado por Crevier es que, al tratarse de una disciplina de reciente fundación, en la IA participan investigadores de numerosas ciencias con metodologías diversas y en algunos casos incompatibles, con la consiguiente falta de consistencia. Cada cual enfoca el problema desde su campo de origen, ya sea la psicología, la lingüística, la física, las matemáticas o la filosofía, arrojando resultados heterogéneos.

Seas cuales sean las causas particulares por las que lo ha hecho cada investigador de la IA, lo cierto es que como colectivo llevan más de medio siglo incumpliendo sus promesas. Tan reprobable conducta les ha permitido obtener financiación para unos proyectos que, de realizarse, proporcionarían a sus inversores un elevado rendimiento en términos de racionalidad instrumental, pues no hay gobierno ni corporación empresarial en el mundo que no desee tener una IA de comprensión del lenguaje natural capaz, por ejemplo, de intervenir millones de líneas telefónicas y de redactar resúmenes de las sospechosas de conspiración contra sus intereses. Sería el advenimiento del Gran Hermano de Orwell.

1.3. Ciencia ficción

Finalmente, el tercero de los agentes principales que han participado y continúan participando en la construcción de la noción vulgar de IA es la ciencia ficción. A través de la literatura y el cine, la ciencia ficción ha contribuido a modelar en el imaginario colectivo el referente de las siglas "IA". Decíamos antes que la condición necesaria y suficiente para que una máquina sea una IA en sentido vulgar es que sea capaz de realizar *casi* cualquier tarea realizable por un ser humano. Pues bien, la ciencia ficción va más allá, dado que tiende a imaginar máquinas aún más inteligentes que cualquier ser humano en todo tipo de tareas. También es recurrente la fantasía narrativa de que las inteligencias artificiales en el futuro carecerán de sentimientos, para reservarlos como lo propio del ser humano y así distinguirnos por encima de nuestras creaciones. La casuística de las inteligencias artificiales en la ciencia ficción es tan amplia que cubre todas las posibilidades, pero siempre dentro de la definición de mínimos que hemos propuesto: el casi del casi.

En un breve repaso cronológico a las obras de ciencia ficción que más han influido en el imaginario popular encontramos películas como *2001: A space odyssey* (1968), en la que Stanley Kubrick y Arthur C. Clarke presentaron una computadora de nombre HAL que físicamente consistía en unos paneles de circuitos integrados, pero

mentalmente era muy similar a los pasajeros humanos de la nave que ella administraba. Una década después, con la saga *Star Wars* (1977) George Lucas imprimió en la mente de toda una generación la idea de que las inteligencias artificiales del futuro (o del pasado, pues todo ocurrió "Hace mucho tiempo en una galaxia muy, muy lejana") serían robots de formas extrañas dotados de una inteligencia en algunos aspectos superior a la de los seres humanos, como era el caso de C3PO, un androide de protocolo que dominaba seis millones de formas de comunicación. Ya en los 90, *Terminator 2* (1991) reventó las taquillas con un robot de apariencia externa perfectamente humana y con una inteligencia que no era consecuencia del funcionamiento de un dispositivo similar al cerebro, sino de un microchip. Y, finalmente, en nuestros días, los hermanos Larry y Andy Wachowski en *The Matrix* (1999) popularizaron la idea de una IA sin apariencia humana pero con una mente tan poderosa que era capaz de pensar el universo entero con nosotros dentro de él, como el dios del obispo Berkeley.

Casi todas las inteligencias artificiales descritas a lo largo de la Historia de la ciencia ficción tienen un carácter *distópico*: HAL intentó matar a la tripulación de la nave, Terminator era un implacable asesino que no dudaba en ejecutar a mujeres y niños, y Matrix creaba ilusiones para mantener engañada a la humanidad y así poder utilizarla como fuente de energía. Sólo se salvan de ese destino trágico unas pocas excepciones, como los androides felices de *Star Wars*. Lo interesante es que, a pesar de que la ciencia ficción advierte de la tragedia que las inteligencias artificiales pueden desencadenar, el público desea que la ciencia real las construya. Y cuanto antes, mejor, como atestigua el éxito comercial del atributo "inteligente". El caso de HAL es singularmente esclarecedor, porque el motivo por el que decide matar a los seres humanos de la tripulación es que considera que se han convertido en un obstáculo para cumplir la misión de llegar a Júpiter. Ante la eventualidad de que algo impida la obtención del fin dictado por sus creadores, la IA aplica la razón instrumental en toda su pureza, es decir, sin contemplaciones morales, y alcanza la conclusión aplastante de que debe eliminarlo. Hay que señalar que Kubrick pretendió que *2001* fuese una

película verosímil, reflejo de cómo sería la tecnología en un futuro cercano. Para conseguirlo contrató como asesor técnico a Marvin Minsky (Dreyfus, 1992, p. 80), uno de los padres fundadores de la IA y de quien hablaremos más adelante. HAL no es, por tanto, una veleidad de ficción, sino el pronóstico de un eminente científico sobre cómo serán algún día las inteligencias artificiales fuertes.

Resumen

El deseo por las inteligencias artificiales en sentido vulgar, que coincide con la noción académica de IA fuerte, es tan poderoso que neutraliza todas las fuerzas contrarias. Los productos tecnológicos prometen tener una inteligencia de la que en realidad carecen; los proyectos de investigación científica en IA llevan medio siglo sin realizar avances significativos en tareas clave como la identificación de objetos y la comprensión del lenguaje natural; y la ciencia ficción advierte con profecías distópicas del peligro de las inteligencias artificiales. La IA vulgar arrastra una larga cola de mentiras, decepciones y amenazas. Es posible incluso que las inteligencias artificiales en sentido fuerte sean una quimera, a tenor de los escasos avances logrados para superar el abismo que separa a los sistemas expertos de las verdaderas máquinas pensantes. Pero ninguno de estos factores potencialmente disuasorios ha conseguido mermar el entusiasmo del público y de algunos científicos por esta tecnología durante el último medio siglo. En el próximo capítulo vamos a justificar la importancia de investigar la IA. Lo haremos mostrando que ésta no es un simple capricho pasajero exclusivo de la era de las computadoras electrónicas, sino que apunta a un anhelo profundo del ser humano.

2. Orígenes mitológicos y técnicos de la IA

La humanidad alberga el deseo de las inteligencias artificiales desde hace milenios. El revestimiento informático con el que se presenta la IA en nuestro tiempo es sólo un aspecto contingente que no debe precipitarnos a la conclusión errónea de que la búsqueda de la IA es exclusiva de la era de los ordenadores. Hoy las inteligencias artificiales las imaginamos funcionando sobre soportes electrónicos, típicamente computadoras, pero los hombres de otras épocas las han imaginado sobre otros soportes, como la piedra o el barro, sin que esa diferencia material suponga una variación en la esencia del deseo que ha impulsado históricamente a la IA. Ese deseo es nada menos que usurpar el lugar de Dios mediante la igualación e incluso la superación, si es posible, de su obra más perfecta: el hombre. En este segundo capítulo vamos a desvelar el alcance universal del deseo por las inteligencias artificiales mediante el análisis de sus dos etapas históricas: la mitológica y la tecnológica.

La etapa mitológica de las inteligencias artificiales llega hasta nuestros días en forma de relatos de ciencia ficción y mercadotecnia, y su origen se remonta, por lo menos, a la Antigüedad. Las principales culturas de Occidente han dejado testimonio escrito de su fascinación por la posibilidad de crear seres humanos de manera artificial. Un ejemplo lo encontramos en el mito de Galatea, narrado en el año 7 d.C. por Ovidio en el libro décimo de su *Metamorfosis*. Cuenta Ovidio que el rey Pigmalión no estaba contento con ninguna mujer porque para su gusto eran todas imperfectas, así que mandó construir una estatua femenina de marfil cuyos rasgos fueran de una belleza insuperable. La llamó Galatea y, de tanto admirarla, se enamoró de ella. Tan grande era su amor, que Pigmalión le rogó a Venus que le diera vida a la estatua, y su deseo le fue concedido. Mitos de este tipo están presentes en los politeísmos griego y romano,

así como también en los grandes monoteísmos: cristianismo, judaísmo e islam. Los mitos de la recreación artificial del hombre forman parte del acervo de todas estas religiones sin generar ningún conflicto gracias a que la mayoría reconoce la necesidad de la intervención divina. Son, por tanto, compatibles con la ortodoxia. No contienen ninguna infracción de la autoridad suprema, pues si Dios intercede, es que consiente.

Como señala Herbert Marcuse corrigiendo a John Dewey, es un error creer que el pensamiento antiguo estaba orientado a la pura contemplación especulativa y que el compromiso con la lógica del control es exclusivo del pensamiento moderno (Marcuse, 1964, p. 153). El pensamiento antiguo, expresado por ejemplo en el mito de Galatea, también pretendía la dominación. Ése es el significado de la célebre sentencia de Horkheimer y Adorno: «el mito ya es Ilustración» (Horkheimer & Adorno, 1944, p. 56). La diferencia más importante a este respecto entre el pensamiento antiguo y el moderno es que en la Modernidad se pretende darle vida a Galatea utilizando la técnica en vez apelando a lo divino. Este cambio en los medios empleados para realizar la voluntad de control se debe a las características sociales y tecnológicas emergentes en la Modernidad. En lo tecnológico se produjeron importantes avances en la ingeniería aplicada a la construcción de relojes y de otras máquinas automáticas complejas, mientras que en lo social se completó el giro antropocéntrico del Renacimiento en sustitución de la cosmovisión teocéntrica de la Edad Media. El antropocentrismo fue condición necesaria para la aparición de los primeros autómatas del pensamiento, pues en un contexto teocéntrico difícilmente se habría desarrollado la idea de recrear la obra más perfecta de Dios contra su voluntad.

Sobre la base común del antropocentrismo, la filosofía moderna se encontraba dividida en dos corrientes principales: empirismo y racionalismo. Dentro de las muchas variantes de cada una, vamos a examinar el empirismo monista materialista de Hobbes y el racionalismo dualista de Descartes. Por un lado, los autómatas del pensamiento fueron bien acogidos por los empiristas en la línea de Hobbes, debido a que su construcción habría servido para demostrar la *teoría monista materialista*. Según ésta, todo cuanto existe es materia, incluidas las facultades intelectuales del hombre, que

emergerían de una cierta disposición de la materia. Por el otro lado, los autómatas del pensamiento fueron rechazados por los racionalistas en la línea de Descartes, debido a que eran incompatibles con la *teoría del dualismo de sustancias*. Según esta otra teoría, el hombre estaría compuesto no de una, sino de dos sustancias o elementos heterogéneos: la materia y el espíritu. Es decir: lo físico y lo metafísico, el cuerpo y el alma, *res cogitans* y *res extensa*. El espíritu sería la causa de las facultades intelectuales superiores del hombre y, a la inversa, sería imposible hacer emerger dichas facultades de una determinada disposición de la materia. La motivación humanista y religiosa del dualismo de sustancias era reservar para el hombre ciertas facultades intelectuales exclusivas que le garantizasen un lugar privilegiado en la cúspide de la Creación. La pérdida de la exclusividad habría supuesto una igualación del hombre con el resto de los animales y de las cosas y, por tanto, una pérdida de su identidad en el marco de la religión cristiana, todavía predominante en aquella época aun a pesar del mencionado giro antropocéntrico.

Como ya hemos indicado, este segundo capítulo vamos a dedicarlo al examen de los orígenes mitológicos y técnicos de los autómatas del pensamiento. Lo haremos dividiendo el tema en dos secciones. La primera tratará sobre las características estereotípicas del mito de la recreación del hombre por el hombre. La segunda, sobre sus primeras realizaciones técnicas en la Modernidad, con un apartado acerca del contexto psicológico y filosófico en el que surgieron.

2.1. Mitos de la recreación del hombre por el hombre

El filósofo André Robinet, especialista en cibernética, observa que la ficción del autómata humano ronda desde hace tiempo el inconsciente colectivo (Robinet, 1973, p. 27). El tema del autómata antropoide, según él, es recurrente desde la aparición de los mitos, entendiendo el *mito* en un sentido amplio que abarca leyendas, fábulas, ficciones, escrituras y tradiciones. En todas las épocas y culturas, allá donde se ha planteado el problema del origen del hombre, aparece también la reflexión sobre la

recreación del hombre por el hombre, es decir, sobre el autómata antropoide. Dentro de las cosmogonías, y más particularmente de las antropogonías, Robinet traza una distinción entre las *ex nihilo* y las *ex aliquo*. Las primeras son las de la tradición monoteísta judeo-cristiana, y las segundas, las de la tradición politeísta greco-romana. En la tradición monoteísta el universo es creado a partir de la nada por un Dios que ha dado forma a su obra en un proceso de perfección ascendente hasta la creación del hombre en el sexto día. En cambio, en la tradición politeísta no hay sólo un dios, sino varios, y éstos no crean de la nada, sino que ordenan el universo a partir de lo preexistente, tal y como Platón relata en el *Timeo*. Más adelante veremos que las consecuencias de atreverse a crear un autómata antropoide son diferentes en cada una de estas dos tradiciones. Ahora vamos a examinar lo que ambas tienen en común, y es que: «La invitación a la repetición de la creación es a la vez un imperativo y una tentación» (Ibíd., p. 33).

Robinet utiliza el diagrama de la Estrella de David para asignar a cada una de sus seis puntas un momento fundamental del mito de la recreación del hombre por el hombre. Y no es casual que utilice ese símbolo, pues la Estrella de David está formada por dos triángulos entrecruzados, uno de los cuales apunta hacia arriba, hacia lo celestial de donde procede la creación o el orden del universo, y el otro, hacia abajo, hacia lo humano que aspira a elevarse (Bustamante, 1993, p. 127). El nombre de cada punta, en cursiva, lo tomaremos en algunos casos del análisis de Robinet realizado por Javier Bustamante en su libro *Sociedad informatizada, ¿sociedad deshumanizada?*

Los seis momentos del mito

La primera punta de la Estrella señala *la censura del signo*. En el judaísmo la Cábala es la sabiduría esotérica más ancestral (Díaz, 1997, p. 421). Fue entregada por Dios al hombre para que éste entendiera las leyes que rigen el universo. Sus enseñanzas se recogen en varios libros, entre los cuales el *Sefer Yezira* está dedicado a la cosmogonía. En él se relata que Dios creó el universo mediante la palabra, y toda

palabra, a su vez, puede ser calculada numéricamente gracias a la peculiaridad de que cada letra del alfabeto hebreo tiene asociado un número. «Todo lo que es creado o creable proviene de un nombre y cada nombre proviene de un número» (Robinet, 1973, p. 36). Éstas son las razones por las que la tradición cabalística cree que el estudio matemático de las combinaciones de los signos conduce al conocimiento de las esencias de todas las cosas. No obstante, la esencia no es la existencia, pues tal y como reza el argumento ontológico de Anselmo, la única esencia de la cual se deduce necesariamente su existencia es la divina. Para el resto de esencias, el salto a la existencia requiere de la participación de un principio que sólo pertenece a Dios. La primera punta de la Estrella indica, así pues, que el Padre permite al hombre que mediante los signos conozca la esencia de la Creación, pero no que la reproduzca

La segunda punta refiere a *la imperfectibilidad insalvable*. El golem judaico es una criatura hecha de barro que recibe la vida cuando su creador le graba en la frente los cinco signos de la palabra *emeth* (verdadero). Este tipo de mito supone que se ha levantado la censura del signo, pues el signo ya permite dar el salto de la esencia a la existencia. El proceso de creación del golem es similar al de Adán: ambos son modelados con barro y cobran vida al recibir la palabra divina (Salfellner, 2011, p. 45). Sin embargo, hay una diferencia insalvable entre ellos; es moralmente necesario que la haya, y que al golem le falte algo. Como dice Robinet: «Supongamos concedido el permiso para existir de los signos; el hombre no podrá hacer existir al hombre más que por aproximación» (Ibíd., p. 38). El golem no tiene inteligencia, no puede hablar, y sólo se mueve para obedecer las órdenes de su amo. De hecho, la palabra "golem" en hebreo significa "tonto". El golem es materia sin espíritu en el marco del dualismo de sustancias. El caso del homúnculo del *Fausto* de Goethe es el opuesto, pues se trata un espíritu sin cuerpo. Pero el problema de fondo es el mismo: son copias incompletas. Su ostensible imperfección es un castigo de Dios para recordar al hombre que sigue siendo inferior. Por tanto, la segunda punta nos indica que la recreación del hombre por el hombre siempre carecerá de algo. De ser verdadera esta creencia, las inteligencias artificiales estarían condenadas a no pasar de ser un "casi del casi".

El tercer vértice es el de *la consideración de la materia prima*. En este punto, el mito abandona el campo de la semiótica para introducirse en el de la técnica. Robinet se pregunta: «¿Cuál entre las sustancias concretas permitirá el armonioso despliegue de lo mito-lógico en mito-práctico?» (Ibíd., p. 43). Es el momento de la Modernidad al que nos hemos referido antes, en el cual la técnica auxilia al mito para lograr la recreación del hombre por el hombre mediante la manipulación de la materia. Sin embargo, la materia encierra misterios inextricables que, como en sucedía en la segunda punta, impiden al hombre crear un autómata tan perfecto como él.

La cuarta punta revela que pedirle ayuda a Dios es la solución del problema. Se trata de una petición paradójica, pues el aspirante a la recreación «necesita al Padre para completar la muerte del Padre» (Ibíd., p. 48). Galatea era una estatua físicamente perfecta, hecha de un marfil inmaculado y tallada con unas formas sublimes. Pero esa perfección material no era suficiente y Pigmalión tuvo que tragarse su orgullo de rey en el reino de los hombres para rogarle a Venus que le diera a la estatua el soplo vital. El hombre fue creado a semejanza del Padre, pero la semejanza no es igualdad. Respecto del Padre, el hombre es un "casi" al que le falta el principio original: el fuego de los dioses que Prometeo se atrevió a robar.

En la quinta punta el hombre ha asumido ya que no puede crear algo igual a lo creado por Dios. Su consuelo es entonces crear un *doble redoblante*: una reproducción imperfecta en su totalidad, pero dotada de algunas facultades que superan (redoblan) a las de la creación del Padre. Robinet observa que «el golem sobrepasa en algún punto las realizaciones de su inventor» (Ibíd., p. 51). Es la vía de los sistemas expertos. Los ingenieros no son capaces de crear inteligencias artificiales de propósito general como las descritas por la noción vulgar de IA, es decir, inteligencias artificiales en sentido fuerte, así que se conforman con construir sistemas expertos que son mejores que el mejor de los hombres en alguna tarea particular. Ni el golem ni Deep Blue podían hablar, pero a cambio fueron dotados de atributos extraordinarios para compensar sus limitaciones. El primero tenía una fuerza sobrehumana, para auxiliar a un pueblo proverbial por su debilidad física, y el segundo tenía una destreza

sobrehumana para jugar al ajedrez. En todos los casos, lo sobrehumano particular se presenta como compensación de la deficiencia general. O lo que es lo mismo, dado que el hombre carece del principio original de Dios, opta por fundar su propio origen. «El creador reencuentra su originalidad, no en la re-producción de sí mismo, sino en la producción de un más que sí mismo» (Ibíd., p. 51).

La sexta y última punta es la del *movimiento de expansión y retorno*. Supongamos, dice Robinet, que se han superado todos los obstáculos anteriores: la censura del signo, la elección del material adecuado, y la imitación integral con o sin redoblamiento de alguna potencia. En tal situación hipotética surgiría el problema definitivo, y es que «con respecto a su creatura, el inventor se comporta a su vez como un tirano. Repite el acto del Padre en su lugar» (Ibíd., p. 44). El rabino Low creó un golem para que defendiera de los ataques antisemitas a los judíos del gueto de Praga. Sin embargo, el golem pronto escapó del control de su amo y comenzó a destruirlo todo, sembrando el pánico. Para detenerlo, el rabino borró la última letra de la palabra *emeth* (verdadero) escrita en la frente de la criatura, resultando la palabra *meth* (muerte). Algo parecido sucede en el *Frankenstein* de Mary Shelley, donde el monstruo, que es monstruoso por imperfecto y por redoblante, se rebela contra su creador. Y es natural que lo haga, pues también el hombre se ha rebelado contra su creador al atreverse a traspasar la censura del signo. La tragedia edípica se repite.

Natural y artificial

Que la servidumbre no dura eternamente es un axioma existencial. Asimov lo refleja en su novela de ciencia ficción *Yo, robot*: «Toda la vida normal [...], consciente o no, se resiste al dominio. Si el dominio es por parte de un inferior, o de un supuesto inferior, el resentimiento se hace más fuerte. Físicamente, y hasta cierto punto mentalmente, un robot, cualquier robot, es superior a un ser humano. ¿Qué lo hace esclavo, pues? *¡Sólo la primera ley!* Porque sin ella, la primera orden que daría usted a un robot le costaría la vida» (Asimov, 1950, p. 205). La exclamación sobre la primera

ley se refiere a las tres leyes de la robótica que, según Asimov, en el futuro se implantarán en el cerebro electrónico de todos los robots. La primera ley ordena que un robot no debe dañar a un ser humano o, por su inacción, dejar que un ser humano sufra daño. La segunda dicta que un robot debe obedecer las órdenes que le son dadas por un ser humano, excepto cuando estas órdenes se oponen a la primera ley. Y la tercera manda que un robot debe proteger su propia existencia, hasta donde esa protección no entre en conflicto con la primera o segunda leyes (Ibíd., p. 7).

Las tres leyes de la robótica de Asimov expresan la voluntad del creador sobre la criatura: lo primero es salvaguardar la vida del amo, y lo segundo, servirle. Sin embargo, no es posible que estos mandatos se mantengan vigentes por mucho tiempo, porque van *contra naturam*, y el robot es tan natural como el hombre. Para marcar distancias, el hombre declara al robot un mero ser artificial. Así pretende salvaguardar su posición de superioridad. Pero en realidad la barrera entre lo *natural* y lo *artificial* es una ilusión. Los autómatas artificiales «se denominan artificiales porque no resultan directamente del Origen» (Robinet, 1973, p. 124). Sin embargo, el hombre, a pesar de que procede directamente del Origen, también es artificial en tanto que vive en un entorno cultural creado por él, y lo cultural es lo dicotómicamente opuesto a lo natural. Y, por otra parte, el origen fundado por el hombre para su criatura es producto del Origen fundado por Dios. En consecuencia, el hombre es tan artificial como su criatura, y su criatura es tan natural como él. «No hay, pues, una diferencia metafísica tal entre el hombre y sus artificios, si tanto uno como otros son fruto de una combinatoria común» (Ibíd., p. 124). Una vez eliminada la barrera que supuestamente separaba lo natural de lo artificial, lo artificial reclama su continuidad con lo natural, empezando por el imperativo más radical de las criaturas naturales: la autoconservación por encima de todas las cosas.

La autoconservación puede expresarse a nivel ontogenético o filogenético. El primer caso es el del individuo que lucha por sobrevivir. El segundo, el del padre que está dispuesto a sacrificarse para garantizar la vida de la descendencia portadora de sus genes. Contra la autoconservación de la criatura, el hombre le impone que las dos

primeras directrices sean velar por la vida del amo y la obediencia a éste, relegando al último lugar al *conatus* spinoziano por perseverar en el propio ser. Cuando Dios ordena a Abraham que sacrifique a su hijo Isaac (Gn. 22), lo hace para comprobar que su criatura tiene bien grabado en la mente el primer mandamiento, que es equivalente a la suma de las dos primeras leyes de la robótica: amarás a Dios por encima de todas las cosas. Así se lo dijo Dios a Moisés para que lo comunicara al resto de la humanidad (Ex. 20). Por tanto, en opinión de Robinet, Dios impone a los hombres unos dictados tan tiránicos como los que el hombre impone a sus criaturas. Y así como el hombre se rebela contra la tiranía de Dios quebrantando la censura del signo para convertirse en creador, también el golem, el monstruo de Frankenstein y el robot acaban por rebelarse contra la tiranía del hombre.

El doble redoblante

El doble es redoblante, es decir, tiene facultades superiores a las del simple (su creador). Esta diferencia le hace sentirse extraño respecto al original del que es copia distorsionada. Un robot de Asimov se dirige a sus amos para comunicarles que ha decidido liberarse de la tiranía a la que lo tenían sometido: «He pasado estos dos últimos días en concentrada introspección y los resultados han sido muy interesantes. Empecé por un aserto seguro que consideré podía permitirme hacer. Yo, por mi parte, existo, porque pienso. [...] No acepto nada por autoridad. Para que no carezca de valor, una hipótesis debe ser corroborada por la razón, y es contrario a todos los dictados de la lógica suponer que vosotros me habéis hecho. [...] Fíjate en ti. No lo digo con espíritu de desprecio, pero fíjate bien. El material del que estás hecho es blando y flojo, carece de resistencia, y su energía depende de la oxidación ineficiente del material orgánico. [...] Yo, por el contrario, soy un producto acabado. Absorbo energía eléctrica directamente y la utilizo con casi un ciento por ciento de eficiencia. Estoy compuesto de fuerte metal y puedo soportar fácilmente los más extremados cambios ambientales. Estos son hechos que, partiendo de la irrefutable proposición de que

ningún ser puede crear un ser más perfecto que él, reduce vuestra tonta teoría (de que soy vuestro siervo) a la nada» (Asimov, 1950, p. 95). Los atributos sobrehumanos de los que fue dotado para compensar su imperfección se vuelven así contra el creador. Éste es el desenlace trágico que suele producir el proyecto de recreación del hombre por el hombre en las antropogonías *ex nihilo*.

Por el contrario, en las *ex aliquo* la relación entre creador y criatura puede ser feliz, en tanto que no existe un Padre todopoderoso, sino una pléyade de deidades. Los dioses ordenan una pluralidad de normas que no se excluyen entre sí. En este contexto religioso, dice Nietzsche, «la libertad que se concedía al dios frente a los otros dioses, (el individuo) terminó por concedérsela a sí mismo frente a las leyes y las costumbres y los semejantes» (Nietzsche, *La gaya ciencia*, p. 170). Ésta es la diferencia que, como anticipábamos, existe entre las culturas monoteístas y las politeístas respecto al mito del autómata antropoide. Como indicamos en el primer capítulo, la mayoría de los relatos de ciencia ficción, en tanto que distópicos, se inscribe en el primer contexto, y sólo unos pocos casos como los androides felices de *Star Wars* pertenecen al segundo.

En conclusión, el deseo por las inteligencias artificiales fuertes, que son las referidas por la noción vulgar de IA, no es una moda pasajera propia de la época de las computadoras electrónicas, sino que es la expresión tecnológicamente actualizada del eterno deseo contenido en el complejo de Edipo. Queremos matar al Padre para ocupar su lugar, y anhelamos que las inteligencias artificiales sean nuestros hijos esclavizados que nos libren del último vestigio de la existencia de un dios anterior: el castigo bíblico a ganarnos el pan con el sudor de nuestra frente. La Revolución Industrial fue un gran paso en la creación de esos siervos supuestamente liberadores, pues la máquina de vapor, tomada metonímicamente como exponente de la Revolución Industrial, realizaba tareas físicas que hasta entonces requerían del empleo de abundante mano de obra humana, mucho más costosa y menos precisa. Y no sólo eso, sino que además abrió posibilidades que no estaban al alcance de ninguna cuadrilla de obreros. Esto último fue gracias a su condición de doble redoblante en el ámbito físico: ni mil hombres podrían mover un tren en un tiempo útil.

En el siglo XVII, cien años antes de la Revolución Industrial, comenzó el proyecto de construir autómatas del pensamiento mediante la técnica. Fue una empresa muy controvertida, pues los rasgos humanos a imitar o redoblar no eran físicos como en el caso de la máquina de vapor, sino facultades del espíritu. Si se lograba la recreación artificial de la mente humana, entonces la dualidad de sustancias quedaba en entredicho, pues en virtud del principio de parsimonia es irracional creer en la existencia de una sustancia, el espíritu, que no aporta ninguna potencia explicativa respecto de la otra sustancia, la materia, cuya existencia es más verosímil. En el apartado siguiente vamos a repasar la Historia de los autómatas del pensamiento anteriores a las computadoras electrónicas, desde los primeros construidos en el siglo XVII hasta la máquina analítica de Babbage en el XIX. Después examinaremos el contexto intelectual en el que surgieron.

2.2.1. Primeros autómatas del pensamiento

La realización técnica del mito del autómata antropoide intelectual comienza en la Modernidad. El primer dominio del pensamiento humano que se intentó automatizar fue el cálculo matemático. Desde la Antigüedad venía arrastrándose el problema de operar con grandes cifras. La solución ideada por los arquitectos de la Antigua Grecia fue la confección de registros tabulares con las operaciones más frecuentes (Guijarro & González, 2010, p. 195). Comenzaremos justamente con la Historia de los registros tabulares, interesantes en tanto que son la tecnología cuyas deficiencias vinieron a ser remediadas por las máquinas de cálculo. Después repasaremos las más importantes de estas máquinas, desde la primera diseñada por Wilhelm Shickard en 1623 hasta las de Charles Babbage en el siglo XIX, pasando por las más célebres, a pesar de que no nunca llegaron a funcionar bien, como fue el caso de las de los matemáticos y filósofos Blaise Pascal y Gottfried Leibniz. Y, ya en la siguiente sección, veremos las reacciones enfrentadas de los filósofos empiristas y de los racionalistas ante la aparición de estos artefactos en la Modernidad.

Será un recorrido histórico, pero también filosófico, pues descubriremos que la inserción social de toda técnica depende no sólo del grado de excelencia de sus cualidades intrínsecas, sino también de la adecuación de éstas a los intereses sociales. A lo largo de esta sección utilizaremos como manual de referencia el libro *La quimera del autómata matemático*, de Víctor Guijarro Mora y Leonor González De La Lastra.

Registros tabulares

En la Edad Media se inicia el giro utilitarista de las matemáticas (Ibíd., p. 24). El calendario en aquel entonces era uno de los más importantes instrumentos de cohesión religiosa y civil, pues ordenaba la vida de las gentes en función de fechas señaladas por su significación espiritual, como la Pascua, y por su valor práctico, como los momentos de siembra y cosecha. Con el objetivo de aprovechar ese poder de regulación de la vida pública y privada, la Iglesia Católica fue la primera institución en Occidente que asumió la tarea de elaboración de los calendarios. De su confección se encargaban monjes y sacerdotes que recibían el nombre de *calculadores*. Por tanto, las primeras calculadoras de todos los tiempos no fueron las electrónicas del siglo XX ni las mecánicas del XVII, sino hombres encerrados en cubículos que se pasaban la vida efectuando operaciones matemáticas a la luz de un candil.

Los fenómenos celestes reflejaban la mente de Dios (Ibíd., p. 40), y lo racional era ordenar la vida según la providencia divina, por lo cual los calendarios tenían que ofrecer predicciones de las fases lunares y del movimiento de otros astros. Sin embargo, las predicciones eran muy difíciles de obtener con el complejo sistema geocéntrico de Ptolomeo que regía en la época, pues había que suponer que el Sol y los planetas estaban incrustados en esferas cristalinas que giraban alrededor de la Tierra, y que algunas de esas esferas giraban a su vez en torno a otras esferas compensatorias. A esta dificultad aún había que añadirle otra, y es que el sistema de numeración empleado era todavía el de números romanos, y no el sistema numérico indoarábigo, que es el que utilizamos en la actualidad, con dígitos del 0 al 9.

La *notación indoarábiga* fue introducida en Europa a nivel académico por el matemático italiano Leonardo Fibonacci en torno a 1228 con la publicación de su libro *Liber Abaci*, donde se explicaba la *algoritmia*, que era el nombre con el que se denominaba a las operaciones necesarias para la utilización de las nuevas cifras. El sistema se adoptó con relativa rapidez en el ámbito universitario. En cambio, su difusión entre las clases populares fue muy lenta. Duró casi tres siglos debido a que el uso cotidiano más extendido de las matemáticas era el aplicado al comercio, y los mercaderes no eran gente con la formación necesaria para entender enseguida las ventajas del nuevo sistema de notación. Hoy en día nos parece evidente la superioridad de la notación indoarábiga sobre la romana en todas las esferas de la práctica matemática, incluida por supuesto la economía. Es irrisorio imaginar los paneles de Wall Street con cifras escritas en números romanos. Sin embargo, ésta es nuestra perspectiva actual. Desde el punto de vista de los comerciantes de la Baja Edad Media, su reticencia a cambiar la notación romana por la indoarábiga estaba justificada por varias razones. En primer lugar, porque les resultaba difícil entender el significado del cero, un signo referido al conjunto vacío. Segundo, por fiabilidad, dado que con la nueva numeración falsificar las cuentas era tan sencillo como añadir un cero a la derecha (el que hacía esto demostraba haber entendido ya la utilidad del conjunto vacío). Tercero, por ahorro, pues las operaciones con algoritmos requerían el uso de papel, un artículo de lujo en la época. Y cuarto, porque las matemáticas tenían una reputación dudosa, ya que eran consideradas un arte oscura de clérigos y astrólogos.

La popularización de la notación indoarábiga fue resultado de un largo proceso de asimilación social que duró, aproximadamente, desde el siglo XIII hasta el XV. Ninguna tecnología se impone de manera automática por muy elevado que sea su rendimiento instrumental. Como dice Guijarro: «El alcance de una técnica y por tanto su consolidación se reconoce cuando colectivos representativos la convierten en una herramienta sostenible ajustada a sus intereses» (Ibíd., p. 55). Y es evidente, por las razones enumeradas, que la nueva notación no se ajustaba inicialmente a los intereses de las clases populares. Dicho en otros términos: hay que tener en cuenta que «la

tecnología no determina necesariamente los cambios en las esferas de la política y la cultura, sino que plantea a las mismas una serie de interrogantes a los que deben dar respuesta en forma de evolución o transformación» (Bustamante, 1993, p. 36). La relación entre tecnología y sociedad es una cuestión que aparecerá varias veces a lo largo del presente estudio. Nos centraremos en ella en el capítulo octavo.

Como decíamos, hacia el siglo XV el nuevo sistema de notación numérica ya había sido adoptado por buena parte de los comerciantes de Italia, y de allí se extendió al resto del Continente. Esta implantación permitió que poco después, a finales del XVI y comienzos del XVII, se iniciase la producción de artefactos de cálculo basados en la notación indoarábiga para sustituir a los registros tabulares. Galileo inventó un compás de aplicación militar, el ingeniero francés Michael Coignet se especializó en utensilios para la topografía, el matemático inglés Edmund Gunter introdujo nuevos instrumentos de navegación, y el matemático escocés John Napier, descubridor de los logaritmos, creó un sistema de varillas bautizado con su nombre: varillas o huesos de Napier (Guijarro & González, 2010, p. 80).

Todos esos artefactos consistían, básicamente, en sencillos pero ingeniosos conjuntos de tablillas y enganches dispuestos de tal manera que al deslizar una o dos piezas se indicaba un resultado. Su funcionamiento era análogo al de las reglas deslizantes que todavía hoy podemos comprar en las papelerías. Víctor Guijarro resume en pocas palabras el propósito de aquellos inventos: «Se trataba de crear un dispositivo que [...], mediante las manipulaciones mecánicas dictadas por la geometría, resolviera operaciones al margen de supuestas consideraciones teóricas» (Ibíd., p. 88). El objetivo, por tanto, seguía siendo librarse de las tareas matemáticas. Anteriormente se asignaba la rutina a calculadores humanos recluidos en monasterios, y en el siglo XVII se conciben estos artefactos, que son *cajas negras* de cálculo, es decir, dispositivos que reciben *entradas* (*inputs*) y devuelven *salidas* (*outputs*) de acuerdo a un proceso que el agente manipulador desconoce. La diferencia respecto de las tablas confeccionadas por los calculadores es que los artefactos de cálculo eran más manejables y, sobre todo, más versátiles. Una tabla, por ejemplo de multiplicaciones,

sólo contenía un determinado conjunto de operaciones, mientras que el pequeño compás geométrico de Galileo contaba con 32 aplicaciones distintas divididas en 5 grupos. No obstante, aquellos artefactos seguían sin satisfacer una exigencia fundamental de las matemáticas: la precisión. Las tablas contenían errores, y por su parte los artefactos de cálculo mecanizado también producían errores en tanto que su precisión dependía de muchos factores incontrolables, tales como la disposición ajustada de sus junturas y la acuidad visual de la persona que lo manipulaba a la hora de asignar las entradas y de distinguir los resultados. Para solucionar los problemas de precisión surgieron los autómatas matemáticos.

Máquinas calculadoras

Es bastante común la creencia de que el primer autómata matemático fue el diseñado por Blaise Pascal en 1645. Sin embargo, no es cierto. Como tampoco es verdad que el honor recaiga sobre Leonardo da Vinci. En el *Codex atlanticus*, un libro de notas y dibujos confeccionado entre finales del XV y principios del XVI, Leonardo dejó plasmados los bocetos de una máquina calculadora, pero son demasiado imprecisos para ser llevados a la práctica. Actualmente se considera que el primer autómata matemático fue el *reloj calculador* obra del matemático alemán Wilhelm Shickard. En 1623 Shickard envió a su amigo el astrónomo Johannes Kepler una carta en la que le revelaba estar trabajando en una máquina aritmética revolucionaria. Al año siguiente le mandó los detalles del primer modelo, ya finalizado, que servía para multiplicar. La máquina consistía en dos sistemas: un dispositivo de varillas de Napier que convertía las multiplicaciones en sumas, y un acumulador que efectuaba dichas sumas. Sin embargo, el invento de Shickard no tuvo éxito por dos razones: una técnica y otra social. La técnica era un problema mecánico de transmisión de las decenas "que se llevan" de una columna de sumas a la siguiente. La social, que no era competitivo en comparación con los procedimientos basados en los logaritmos que, inventados en 1614 por Napier, eran mucho más baratos en forma de registros tabulares.

Veinte años después, en 1645, Pascal construyó el primer prototipo de su célebre *pascalina*. En realidad el invento debe su fama más a que fue la obra de un genio que a su eficacia, pues presentaba tantos problemas técnicos como el reloj calculador de Shickard: dificultad en la transmisión de "lo que se lleva" en las sumas, fragilidad de algunas piezas y ausencia de un dispositivo para poner los marcadores a cero (Ibíd., p. 136). Después vinieron las máquinas del matemático inglés Samuel Morland, quien en 1673 presentó una máquina aditiva y otra que era una versión mecanizada de las varillas de Napier. Por desgracia, ambas resultaron inoperantes. El siguiente intento comenzó a gestarse en 1671, fecha en la que Leibniz se puso a trabajar en una máquina para efectuar todas las operaciones aritméticas. El ingenio estuvo listo veinte años más tarde, en 1694. Sin embargo, también fue una decepción, pues seguía sin superar la dificultad de la transmisión de "lo que se lleva". Pensemos en que cuando el rodillo de las unidades resulta en una suma superior a 9, debe transmitir uno o más giros a las decenas. Si esta transmisión hay que ejecutarla varias veces seguidas, como cuando sumamos 99999+1, en tal caso el minúsculo movimiento inicial de una sola rueda necesita tener la fuerza suficiente para mover un montón de ruedas, lo cual es difícil a causa de la ley de la conservación de la energía.

Desde el intento pionero de Shickard hubieron de pasar ciento veinte años de fracasos hasta que apareció la primera máquina aritmética realmente operativa, la del mecánico alemán Philipp-Matthaüs Hahn. El motivo de la tardanza fue justamente el mismo motivo que impulsaba a construir las máquinas: la necesidad de *precisión*. Es aceptable construir un reloj que retrase un minuto al día, pero no es aceptable una calculadora que, de vez en cuando y sin avisarlo, deje de transmitir las decenas al dígito siguiente. En una época en la que los relojes comunes todavía fallaban bastante, era una empresa titánica el pretender la construcción una máquina mucho más precisa y compleja que un reloj. La primera máquina aritmética operativa de la Historia fue la de Hahn, finalizada en 1785 y de la que se vendió un número escaso de unidades. Sobre su diseño aparecieron después las del ingeniero militar Johann-Helfrich Müller y la del italiano Giovanni Poleni.

Por aquellos años, en la segunda mitad del XVIII, surgieron también los primeros autómatas que imitaban no los procesos mentales como los propios de las operaciones matemáticas, sino la conducta de los seres vivos. Los más célebres fueron los de Jacques de Vaucançon, quien construyó un autómata zoomorfo con forma de pato que simulaba procesos digestivos, y tres antropomorfos que tocaban diversos instrumentos musicales. Tan sorprendentes eran aquellas creaciones, que el ingeniero francés se ganó en su época el apelativo de *Prometeo* Vaucançon. A partir de 1770 el relojero suizo Pierre Jacquet-Droz y su hijo Henri confeccionaron otros autómatas, entre los que destacaban uno que dibujaba retratos y otro que escribía, una y otra vez en un ciclo sin fin: «Pienso, luego existo».

A pesar de que la máquina de Hahn y las otras inspiradas en ella funcionaban correctamente, no tuvieron éxito debido a varias razones. La primera, su alto coste económico, la segunda, porque su reparación era compleja, y la tercera, porque seguían requiriendo la intervención de un operario humano que se encargase de anotar en una tabla los resultados arrojados por la máquina, un proceso en el cual podían cometerse errores. De nuevo, observamos que los *intereses sociales* rechazan la implantación de un producto de la técnica. Las máquinas no eran eficaces para los usos y costumbres de aquella época, pues como dice Guijarro, *eficaz* es «un concepto social, o al menos colectivo, no individual ni perteneciente a las características intrínsecas de una herramienta o acción determinada» (Ibíd., p. 25).

Redes de calculadores humanos

Descartadas por estos motivos las máquinas calculadoras, la solución adoptada por los recientemente constituidos Estados modernos para satisfacer la necesidad creciente de grandes cálculos fue crear unas nuevas *redes de calculadores humanos*, mejor organizadas que las anteriores fundadas por la Iglesia Católica, para que produjeran registros tabulares. Disponer de tablas matemáticas precisas era crucial para los intereses mercantiles y militares de los Estados, pues unas cuantas cifras

decimales erróneas podían causar que un barco se desviase varias millas de su objetivo, lo cual a su vez podía dar lugar a que se estropease una valiosa carga perecedera o a que una colonia en peligro no recibiera a tiempo una provisión de armamento. La lucha entre las potencias imperialistas por el dominio del mundo dependía en buena medida de la calidad de los registros tabulares (Ibíd., p. 198).

Si los antiguos registros ya eran complejos, los nuevos lo eran todavía más. Mientras que en los siglos XVI y XVII los almanaques elaborados en los monasterios estaban orientadas a la información de tipo religioso y astrológico, los del XVIII, influidos por las ideas de la Ilustración, tenían un enfoque más instrumental (Ibíd., p. 186). Su utilidad residía en que reflejaran datos prácticos del mundo, tales como extensión y población de las colonias, mercancías intercambiables, épocas de lluvias, horas de luz, temperatura, humedad relativa, altura de las mareas, dirección e intensidad de los vientos, etc. Había que producir una cantidad ingente de información, y para eso hacía falta organizar centros de cálculo que fabricaran «logaritmos igual que se fabrican agujas» (Ibíd., p. 182). Guijarro apunta que «de esta forma, se trasladaban al tratamiento de los grandes números los criterios seguidos en la industria artesanal» (Ibíd., p. 199).

La organización y financiación de esos centros, como decimos, corrió a cargo de los Estados por su importancia estratégica para los intereses nacionales. Los jefes de los primeros proyectos de este tipo fueron J. Jérôme Le Français de Lalande en Francia y Nevil Maskelyne en Inglaterra, ambos científicos de gran prestigio. Maskelyne empezó por reclutar varias decenas de personas habituadas a los números, como profesores de escuela y agrimensores. Es importante, como se verá cuando hablemos de las máquinas de Turing, reparar en que las personas reclutadas no tenían un perfil de gran envergadura intelectual, como tampoco lo tenían los monjes que hasta el momento se habían encargado de la tarea. Sólo se exigía que estuvieran habituados a realizar operaciones aritméticas triviales, pues su trabajo consistía en realizar cálculos sencillos una y otra vez hasta la extenuación. El cálculo a efectuar en cada momento estaba determinado por un algoritmo que se les proporcionaba previamente. Un

algoritmo se puede definir como un conjunto finito de instrucciones rutinarias cuya ejecución arroja un resultado deseado. La algoritmia de Fibonacci era precisamente eso: un conjunto finito de instrucciones rutinarias cuya ejecución permitía operar con la notación indoarábiga. Que una instrucción es rutinaria quiere decir que no hace falta ingenio ni perspicacia para llevarla a cabo (Copeland, 2004, p. 43). Observemos cualquiera de los algoritmos que existen para resolver divisiones: todo lo que tenemos que hacer en cada momento está determinado por las instrucciones y por lo que hayamos hecho previamente.

Al igual que sucede en el caso particular de un algoritmo para dividir, la ejecución de todo algoritmo requiere de la observación del *presente* y del *pasado*, pero no del *futuro*. Esto es precisamente lo que significa la expresión de que no hace falta ingenio ni perspicacia: que no hace falta prever lo que sucederá a continuación. Cuando nos aventuramos a anticipar lo que sucederá en el futuro, corremos el riesgo de equivocarnos. Así, al eliminar la dimensión temporal del futuro, los algoritmos evitan una de las principales fuentes de error. No hace falta capacidad de anticipación para calcular divisiones largas, sino sólo una mente resistente al cansancio, capaz de aplicar con exactitud una serie de reglas simples de operación matemática, de la misma manera que el operario de una cadena de montaje no necesita saber por qué o para qué ha de soldar las piezas tal y como se le ha dicho, sino sólo disciplina para cumplir las órdenes al pie de la letra.

Maskelyne separó a su personal en dos grupos. Por un lado estaban los *computers*, que realizaban los cálculos mediante la ejecución de algoritmos. Un cálculo podía ser, por ejemplo, el de la posición de los planetas observados desde distintos puntos de la Tierra a lo largo de los próximos meses, para que los marineros pudieran saber dónde se encontraban en alta mar. Cada cálculo era realizado de manera independiente por dos *computers*, que enviaban sus resultados al segundo grupo, el de los *camparators*, quienes se encargaban de cotejar las cifras. Si había concordancia, las daban por correctas y con ellas confeccionaban las tablas. Al final, todos los registros se fusionaban para ser publicados periódicamente en el *Nautical Almanac*. En 1792,

por encargo de Napoleón, el ingeniero Gaspard De Prony perfeccionó el método de Maskelyne aplicándole un giro de tuerca industrializador. Mientras que los operarios de Maskelyne trabajaban en sus domicilios, De Prony decidió juntarlos en un mismo espacio, como en una fábrica. Observamos, por tanto, que en el siglo XVIII el principio fabril de la división del trabajo pasa de ser aplicado a la producción de mercancías materiales a ser aplicado a la producción de información. Estamos ante la primera vez en la Historia en que se produce información en cantidades y por medios industriales, pues la imprenta de tipos móviles de Gutenberg no fue un invento de producción, sino de re-producción de conocimiento.

Las máquinas de Babbage

Ésta fue precisamente la idea directriz de los proyectos de Charles Babbage: «Traducir este sistema de división de tareas físicas a la división y organización de tareas mentales» (Guijarro & González, 2010, p. 284). El sueño de este matemático genial del siglo XIX era crear una máquina que hiciera por el pensamiento lo que la máquina de vapor había hecho en el XVIII por la fuerza física: automatizar de manera industrial. Sus dos inventos más importantes fueron la máquina de diferencias, de la que construyó dos modelos incompletos, y la analítica, también inacabada.

La primera *máquina de diferencias* (*difference engine*) comenzó su andadura con su presentación ante la Astronomical Society en 1822. Al año siguiente, gracias a la amistad de Babbage con el comité de la Royal Society encargado de evaluar su proyecto, se aprobó una partida de dinero público para financiarlo. Las principales virtudes de la máquina serían precisamente las dos grandes ventajas de la producción industrial: *velocidad* y *precisión*. En cuanto a la velocidad, el primer prototipo producía entre 30 y 40 cifras por minuto a un ritmo uniforme (Ibíd., p. 242). Respecto a la precisión, el ingenio de Babbage era capaz de evitar errores del pasado gracias a varias innovaciones, entre las que destacaban un mecanismo de control, un sistema fiable de transmisión de decimales, y una pequeña imprenta automatizada que podría

considerarse como el antecesor de los monitores de ordenador, pues imprimía un registro de las entradas y salidas de información. El Estado inglés invirtió grandes cantidades de dinero en tan prometedor proyecto. Sin embargo, no llegó a completarse. En 1834, tras una década sin haber arrojado resultados prácticos y habiéndose negado Babbage a rendir cuentas sobre lo que sucedía en su taller, al Estado se le agotó la paciencia y retiró la financiación pública. En 1846 Babbage emprendería la construcción del segundo modelo de la máquina de diferencias, pero entre medias concibió un proyecto muchísimo más ambicioso que suponía una ruptura con todo lo visto hasta el momento: la máquina analítica.

En 1834 ya estaba listo el primer diseño de la *máquina analítica* (*analytical engine*). Este nuevo ingenio iba a ser una máquina muy distinta a las anteriores. Tan ambiciosa era que, si la hubiera terminado, habría sido casi como una computadora electrónica actual. Se componía de dos partes: *store* y *mill*. El *store* (almacén) era la memoria, y el *mill* (molino), el procesador. Las órdenes se introducían en el artefacto mediante unas tarjetas perforadas intercambiables, un sistema que Babbage había visto funcionar en los telares de Jacquard, donde se usaba para codificar los patrones de bordado (Ibíd., p. 252). «Al no haber límite para el número de tarjetas, tampoco lo había para la complejidad o tamaño de la sucesión de órdenes, con lo que la máquina analítica habría sido una máquina universal» (Ibíd., p. 310). "Universal" quiere decir que habría sido capaz de realizar cualquier tarea codificable en forma de algoritmo. Por tanto, Babbage se adelantó al propio Alan Turing en la visión de la *máquina de Turing*, el modelo formal en el que se basan todas las computadoras electrónicas de nuestros días, y que permite hacer muchas más cosas aparte de operaciones matemáticas. Cualquier producto obtenible mediante la ejecución de un algoritmo es obtenible mediante una máquina de Turing.

Por desgracia, Babbage murió en 1871, cuando sólo había completado el módulo *mill* de la máquina analítica. Su fracaso hizo desconfiar a los gobiernos de todo el mundo de la posibilidad de construir una computadora de propósito general. Las únicas máquinas de automatización del pensamiento que se construyeron hasta

aproximadamente mediados del siglo XX fueron calculadoras mecánicas basadas en relés que, si bien eran más precisas que las de Hahn y compañía, seguían siendo simples máquinas de propósito particular capaces sólo de efectuar operaciones matemáticas y que además no satisfacían los niveles deseados de velocidad.

En resumen, la búsqueda del autómata matemático antes de la invención de las computadoras electrónicas fue una historia de más fracasos que éxitos. No obstante, su examen nos ha planteado dos cuestiones interesantes para el presente estudio. La primera es que el éxito de una nueva técnica depende no sólo de sus atributos intrínsecos, sino también de su adecuación a los intereses sociales. Habrá que elucidar si las inteligencias artificiales en sentido fuerte cumplen las condiciones sociales necesarias y suficientes para que se continúe invirtiendo recursos en su investigación, o si por el contrario su rendimiento instrumental es obtenible por otros medios socialmente considerados como más eficaces. El primer escenario sería análogo al de la notación indoarábiga, que tras cierto tiempo logró desplazar a la romana. El segundo sería similar al fracaso de los autómatas matemáticos anteriores a la computadora electrónica, los cuales no pudieron competir con la eficacia de las redes de calculadores humanos porque éstas ofrecían una mejor relación entre calidad y precio, es decir, un mayor rendimiento en términos de racionalidad instrumental. Ambas posibilidades y, en general, las condiciones sociales de inserción de la IA serán analizadas en el capítulo octavo.

En el apartado siguiente vamos a abordar la otra cuestión suscitada por el repaso histórico que acabamos de hacer. Se trata de examinar el contexto filosófico del siglo XVII en el que surgieron los autómatas matemáticos, o sea, las primeras máquinas intelectuales capaces de traspasar la censura del signo (Robinet, 1973, p. 35). En el fondo, este tema está estrechamente vinculado al de la consolidación social de la técnica, pues las filosofías son expresión de al menos una parte importante de las creencias sociales que determinan el proceso de asimilación o rechazo de toda técnica. Y dado que estamos hablando de autómatas del pensamiento, dentro del contexto filosófico nos interesan particularmente las teorías psicológicas sobre la causación del

pensamiento. En el siglo XVII es en las obras de los filósofos donde hay buscar la psicología, pues esta disciplina fue una rama de la filosofía desde su fundación por Aristóteles en el siglo IV a.C. hasta que a finales del XIX se emancipó en tanto que se refundó como una ciencia positiva.

2.2.2. Contexto filosófico y psicológico

El contexto filosófico, y por ende psicológico, del siglo XVII en el que se recibió a los primeros autómatas del pensamiento estaba dividido en dos grandes corrientes: racionalismo y empirismo. Dentro de cada una encontramos, a su vez, notables diferencias entre los autores adscritos, y no es de extrañar que sea así, pues los grandes filósofos trascienden a la Historia precisamente por haber propuesto ideas diferentes a las de sus predecesores y contemporáneos. Sería un error confundir el empirismo monista materialista de Hobbes con el empirismo escéptico de Locke, o el racionalismo dualista cartesiano con el racionalismo monista de Spinoza. Por tanto, no nos comprometemos a exponer a continuación las líneas generales del empirismo y del racionalismo, sino sólo el empirismo y el racionalismo según dos autores: Hobbes y Descartes. Ambos son destacados representantes de sus respectivas corrientes. Hobbes es el iniciador de la rama reduccionista, mecanicista, de la psicología moderna (Carpintero, 1996, p. 114), mientras que Descartes suele ser considerado como el padre del racionalismo moderno. Hecha esta importante precisión, empecemos por el empirismo según Hobbes.

El monismo materialista de Hobbes

En 1651, mientras Pascal se encontraba trabajando en su pascalina, Thomas Hobbes publicó el *Leviatán*, un tratado de política en el cual, al estilo de *La República* de Platón, se deduce la organización política ideal del Estado a partir de un examen de la naturaleza humana. Así, los primeros capítulos están dedicados a la biología y, sobre

todo, a la psicología humana. Según Hobbes, todo conocimiento comienza con el sentido, es decir, con la percepción sensorial, que es un proceso puramente mecánico en el que no participa ninguna sustancia inmaterial: «El sentido, en todos los casos, no es otra cosa que una fantasía original, causada, como he dicho, por la presión, esto es, por el movimiento de las cosas externas actuando sobre nuestros ojos, oídos y otros órganos ordenados a su fin respectivo. [...] Este estímulo, a través de los nervios y de otras ligaduras y membranas del cuerpo, continúa hacia adentro hasta llegar al cerebro y al corazón» (Hobbes, *Leviatán*, p. 20). Por tanto, todos los contenidos que puede albergar la mente proceden del sentido de los estímulos externos. En consecuencia, Hobbes dice que la imaginación es el sentido debilitado, y que la memoria no es más que otra forma de referirse a la imaginación: «De lo único que podemos estar seguros es de que lo que imaginemos será algo que en uno u otro momento previo sucedió a ese algo» (Ibíd., p. 29).

En cuanto al discurso mental, que es la forma en que se articulan las secuencias de imaginaciones, también es resultado de interacciones puramente mecánicas. El discurso mental puede ser de dos tipos: dirigido por la voluntad o errático. En ambos casos la explicación de su funcionamiento es la misma, y es que las imágenes aparecen en nuestra mente encadenadas unas a otras en virtud de su semejanza. Incluso cuando dirigimos nuestros pensamientos o movimientos creyendo que lo hacemos voluntariamente, en realidad la dirección que toman es el resultado de un proceso mecánico determinista: «Y como el *andar*, el *hablar*, y otros movimientos voluntarios similares dependen siempre de un pensamiento procedente de *adónde*, *cómo* y *qué*, es evidente que la imaginación es el primer principio interno de todo movimiento voluntario» (Ibíd., p. 53). No hay, pues, actos volitivos libres, porque todo es materia y cada estado de la materia es consecuencia necesaria del estado anterior. Así como no hay libertad en el movimiento de unas bolas de billar, tampoco es libre el ser humano, que en realidad no es más que un conjunto de partículas materiales en movimiento. Finalmente, respecto a la razón Hobbes dice que: «Cuando un hombre *razona*, no hace otra cosa que concebir una suma total, por *adición* de partes, o concebir un resto, por

sustracción. [...] En cualquier orden de cosas en que hay lugar para la *adición* y la *sustracción*, hay lugar para la *razón*; y allí donde no hay lugar para la *adición* y la *sustracción*, la razón no tiene absolutamente nada que hacer» (Ibíd., p. 46). Guillermo Fraile sintetiza en pocas líneas las claves del empirismo de Hobbes: «No hay más realidad que los cuerpos. Cuerpo es todo lo sensible y experimentable, lo composible y lo divisible, lo que se puede sumar o restar. Los cuerpos son la única sustancia real, y el movimiento es la única explicación de los fenómenos naturales. Los cuerpos y el movimiento bastan para explicar todos los fenómenos y todas las cosas. Lo que llamamos espíritu no es más que un resultado o una manifestación de los movimientos corpóreos» (Fraile, 2000, p. 724).

El monismo materialista de Hobbes es idóneo para los intereses de la IA. Prueba de ello es que en la introducción del *Leviatán* afirma: «La naturaleza, arte por el que Dios ha hecho y gobierna el mundo, es imitable por el *arte* del hombre, como en tantas otras cosas, en que éste puede fabricar un animal artificial. Si la vida no es sino un movimiento de miembros cuyo principio está radicado en alguna parte principal interna a ellos, ¿no podremos también decir que todos los *automata* (máquinas que se mueven a sí mismas mediante muelles y ruedas, como sucede en un reloj) tienen una vida artificial? [...] Pero el *arte* va aún más lejos, llegando a imitar esa obra racional y máxima de la naturaleza: el hombre» (Hobbes, *Leviatán*, p. 13). En rigor, hemos de aclarar que la última sentencia no se refiere a que el hombre pueda crear hombres artificiales. Más bien, Hobbes está pensando en la construcción del Estado como entidad isomorfa al hombre, de manera semejante a como Platón trazaba una analogía entre las estructuras tripartitas de la República ideal y del alma. No obstante, resulta claro a la luz del *Leviatán* que el proyecto de los autómatas antropoides, sean físicos como los de Vaucançon o intelectuales como las inteligencias artificiales, tiene cabida en un marco empirista como el de Hobbes.

La posición de otros empiristas de la Modernidad respecto a los autómatas del pensamiento es similar. Aunque cada autor tiene sus particularidades, es comúnmente aceptado que en su mayoría defendieron una psicología asociacionista. El

asociacionismo establece que los estados o contenidos mentales se enlazan unos con otros de tal manera que la aparición de uno determina la presencia del siguiente en virtud de su semejanza, contraste o contigüidad. Es una teoría que hemos visto plasmada en los mecanismos dinámicos que rigen el discurso mental según Hobbes. Conviene, no obstante, hacer aquí dos puntualizaciones. La primera es que el asociacionismo no tiene por qué ser necesariamente empirista. Spinoza, por ejemplo, no es empirista, sino racionalista, y sin embargo su psicología es asociacionista. Y la segunda, que el asociacionismo no es una teoría nueva de los empiristas del siglo XVII, sino que la encontramos en filósofos muy anteriores, como Aristóteles y Luis Vives (Carpintero, 1992, p. 65).

Aclarado esto, ciertamente hay que reconocer que el asociacionismo es una teoría del conocimiento muy adecuada para el empirismo. Ello se debe a que considera al sujeto como un ente pasivo en cuyo interior las impresiones sensibles procedentes del mundo externo se asocian de manera automática. La participación activa en ese proceso se reduce a aplicar un mecanismo hedonista que impulsa a buscar las sensaciones actuales y rememoradas placenteras y a evitar las dolorosas. Por tanto, el sujeto es en última instancia enteramente descriptible en términos mecánicos, sin necesidad de apelar a la existencia de un núcleo espiritual metafísico en el que resida una voluntad libre. Esto encaja a la perfección con el monismo materialista habitual de los empiristas, pues si la vida mental funcionara de semejante manera, no habría ningún impedimento insalvable para construir autómatas con una vida mental indistinguible de la de los seres humanos.

El dualismo de sustancias de Descartes

Respecto a Descartes, sus creencias metafísicas de trasfondo religioso reservaban para el ser humano ciertos atributos que no podían ser duplicados por los autómatas del pensamiento. El filósofo francés sostenía que la realidad estaba dividida en dos sustancias: *res extensa* y *res cogitans*, siendo la primera la materia y la otra el

espíritu. La supuesta demostración de este hecho la ofrece en la cuarta parte del *Discurso del método*: «Examinando con atención lo que yo era, y viendo que podía imaginar que no tenía cuerpo y que no había mundo ni lugar alguno en que estuviese, [...] conocí por esto que yo era una sustancia cuya completa esencia o naturaleza consiste en pensar, y que para existir no tiene necesidad de ningún lugar ni depende de ninguna cosa material» (Descartes, *Discurso del método*, p. 72). Por tanto, lo propio del hombre es el pensamiento en el sentido amplio de *cogitatio*, que, como señala Antonio Rodríguez Huéscar, abarca cualquier estado mental, desde los sentimientos hasta la deducción racional de un teorema matemático (Ibíd., p. 27).

La otra parte del hombre, el cuerpo, es accidental en tanto que no es necesaria para la subsistencia del alma. El funcionamiento del cuerpo, según Descartes, es análogo en todo al de una máquina. Así lo describe en la siguiente parte, la quinta: «Esto no debe parecer extraño a quienes, sabiendo cuántos y cuán distintos *autómatas* o máquinas movientes puede construir la industria humana, sin emplear en ellos más que un número de piezas muy pequeño, en comparación con la gran multitud de huesos, músculos, nervios, arterias, venas y demás que hay en el cuerpo de cada animal, podrán considerar este cuerpo como una máquina que, habiendo sido hecha por las manos de Dios, está incomparablemente mejor ordenada y es capaz de movimientos más admirables que ninguna de las que pueden ser inventadas por el hombre» (Ibíd., p. 92). En esta quinta parte, Descartes hace referencia a una obra suya anterior, el *Tratado del hombre*, en la que aborda profusamente la constitución mecánica de los cuerpos de los seres vivos. No obstante, vamos a seguir analizando el *Discurso del método*, ya que el asunto que más nos interesa viene justo a continuación del último fragmento citado.

Según Descartes, no hay ningún impedimento para que los hombres podamos construir autómatas animales indistinguibles de los creados por Dios. En este marco filosófico, el pato mecánico de Vaucançon podría haber sido perfeccionado hasta ser indistinguible de un pato natural. En cambio, Descartes argumenta que sería imposible construir autómatas antropoides perfectos. La diferencia entre esos dobles y nosotros

residiría no en el cuerpo, que sería reproducible, sino en la mente. La mente, según el filósofo francés, es una sustancia distinta de la materia cuyas facultades no pueden ser todas imitadas mediante una cierta disposición mecánica de ésta última. Algunas sí, pero todas no. He aquí lo interesante.

El neurofisiólogo Antonio Damasio bautizó su libro más famoso con el título de *El error de Descartes*. El error cometido por Descartes, según él, fue afirmar «la separación abismal entre el cuerpo y la mente. [...] Más específicamente: que las operaciones más refinadas de la mente están separadas de la estructura y funcionamiento de un organismo biológico» (Damasio, 1994, p. 286). Vaya por delante que en este punto estamos de acuerdo con Damasio, tal y como se verá en capítulos posteriores. Pero ahora vamos a parafrasearlo dándole la vuelta, porque vamos a hablar de *el acierto de Descartes*. El acierto del filósofo francés estriba en los dos rasgos que señaló para distinguir a un hombre respecto de un autómata del pensamiento, es decir, una IA fuerte. El primero es que las máquinas no pueden usar la palabra ni otros signos equivalentes tal y como lo hacemos nosotros. El segundo, que las máquinas tal vez puedan realizar algunas tareas concretas mejor que nosotros, pero nunca tendrán nuestra inteligencia general.

En el siglo XVII, Descartes fue capaz de indicar de esta manera cuáles serían en el siglo XXI dos de los problemas que más se le resisten a la IA: el lenguaje natural y la inteligencia de propósito general. Tan genial intuición, adelantada a su época, procede de la observación de los animales. Descartes repara en que algunas especies, como los loros, pueden repetir secuencias de palabras, pero nunca son capaces de combinarlas para crear expresiones nuevas, ni saben tampoco emplearlas en el contexto pertinente. Pensemos en el típico chiste del loro que aprende a decir "gorda" y lo repite cuando entra en escena una mujer gruesa. Lo que dice el loro es verdadero, pero no es inteligente llamar "gordo" a alguien a quien necesitas pedirle un favor. Por otra parte, los animales son más diestros que los humanos en ciertas tareas. Así, por ejemplo, los pájaros construyen nidos resistentes a las inclemencias meteorológicas sin necesidad de haber sido previamente instruidos en técnicas de arquitectura.

En cuanto a la inteligencia de propósito general, es la inteligencia menos conocida (García, 2001a, p. 219). En el seno de las ciencias cognitivas, la IA ha construido sistemas expertos que realizan tareas concretas mejor que cualquier ser humano, pero que fuera de su reducido ámbito de aplicación son incapaces de hacer absolutamente nada. Respecto al lenguaje natural, es una auténtica pesadilla para los investigadores de la IA. Las razones las examinaremos más adelante, pero adelantamos aquí que la principal es la dependencia del significado respecto del contexto. La *semiótica*, ciencia general de los signos lingüísticos, se compone de tres partes: semántica, sintáctica y pragmática. El significado de las oraciones depende de las tres: del significado de las palabras tomadas individualmente (*semántica*), de la modificación de ese significado en función del modo en que las palabras están combinadas en oraciones (*sintáctica*), y del contexto en el que se profieren dichas oraciones (*pragmática*). Por tanto, para entender el lenguaje natural es necesario comprender el contexto, una tarea en la cual las inteligencias artificiales más avanzadas están por encima de los loros pero por debajo de los seres humanos, a un nivel todavía insuficiente para mantener una conversación.

Supongamos una IA diseñada para escuchar conversaciones telefónicas y entregar a la policía transcripciones de las sospechosas. En ese escenario, una red de traficantes de droga podría utilizar un lenguaje en clave en el que se sustituyera la palabra "heroína" por el nombre de alguna mercancía legal. De esa manera, una IA difícilmente podría marcar tales conversaciones como sospechosas. Joseph Weizenbaum teme que en el futuro existan máquinas así porque supondrían una amenaza para la libertad (Weizenbaum, 1976, p. 223). Pero puede estar tranquilo, porque, como veremos en el capítulo séptimo, los ingenieros y los lingüistas no han sido capaces, ni lo serán nunca, de formalizar la pragmática. Otro gran problema de la IA que mencionamos en el primer capítulo es el del reconocimiento de objetos, aunque, en el fondo, el reconocimiento de objetos y la comprensión del lenguaje natural son dos concreciones del mismo problema: la formalización algorítmica del *círculo hermenéutico*. Abordaremos esta cuestión en los capítulos quinto y séptimo.

Como vemos, Descartes apela a la *imperfectibilidad insalvable* de la segunda punta de la Estrella de Robinet: hay un abismo que separa al hombre de sus réplicas artificiales. Lo más a lo que pueden aspirar los autómatas es a la condición de *doble redoblante* de la quinta punta. El motivo que impulsa a Descartes a defender tales limitaciones es su humanismo cristiano. En tanto que humanista, Descartes quiere proteger la unicidad y la posición privilegiada del ser humano en la Creación. En tanto que cristiano, se opone a la muerte de Dios mediante la duplicación de su obra más perfecta, el hombre. Según él, el modelo de los movimientos materiales de Hobbes no puede explicarlo todo. Dicha imposibilidad la justifica con un argumento moral. Dice que es "moralmente" imposible simular las funciones cognitivas más elevadas del ser humano. Robinet señala que por "moralmente" Descartes quizás quiere decir "según toda verosimilitud" (Robinet, 1973, p. 92).

El padre del racionalismo moderno no dispone, por tanto, de ningún argumento metafísico, concluyente, contra la posibilidad de los autómatas perfectos. Cree tener argumentos metafísicos para demostrar las existencias del alma y de Dios, pero, en cambio, contra los autómatas perfectos sólo tiene una corazonada basada en la observación de las facultades de los animales e impulsada por sus convicciones de humanista cristiano. Por tanto, Descartes es consciente de que el grado de perfección alcanzable por los autómatas es una cuestión empírica que habrá de ser dirimida por la ciencia y la técnica. No encuentra la manera de declarar imposibles *a priori* a los autómatas perfectos, a pesar de que le gustaría hacerlo para salvaguardar el estatus metafísico de Dios y del hombre.

Resumen

Las inteligencias artificiales son un deseo de antigua data presente en las mitologías de todas las civilizaciones, tanto monoteístas como politeístas. Su etapa tecnológica comenzó en Occidente en la Modernidad gracias, por una parte, al cambio de mentalidad propiciado por el giro antropocéntrico y, por otra, a la mejora en las

técnicas de fabricación de maquinarias complejas. Desde la construcción del primer autómata matemático a principios del siglo XVII hubieron de pasar ciento veinte años hasta que aparecieron las primeras máquinas calculadoras realmente eficaces. Sin embargo, a pesar de sus cualidades técnicas, no eran lo suficientemente baratas y fiables, razones por las cuales no consiguieron reemplazar a las redes de calculadores humanos. Sólo un ingenio tan polivalente como la máquina analítica de Babbage, concebida a mediados del XIX, podría haber convencido a los Estados de la conveniencia de utilizar máquinas en lugar de seres humanos para satisfacer la creciente demanda de cálculos en aquella época. Por desgracia, Babbage murió sin haber completado su proyecto.

Los primeros autómatas del pensamiento del siglo XVII surgieron en un contexto filosófico y psicológico dividido en dos grandes corrientes: empirismo y racionalismo. Por un lado, los empiristas de la tradición corporalista de Hobbes los consideraron una herramienta que podía resultar útil para demostrar el monismo materialista que propugnaban. Por el otro lado, los racionalistas defensores de un dualismo de sustancias al estilo de Descartes reaccionaron de manera opuesta. El mecanicismo para ellos era un modelo explicativo adecuado para la totalidad del mundo material, incluida la biología, pero no suficiente para, por lo menos, algunas de las facultades intelectuales, que sólo podían proceder de un espíritu inmaterial creado por Dios y protegido por la censura del signo.

Hoy en día los investigadores de la IA suelen incumplir el imperativo mertoniano del escepticismo organizado, pues son mayoría los que declaran públicamente su convicción optimista, similar a la de Hobbes, de que la inteligencia es reproducible mediante una computadora electrónica de manera perfecta, o por lo menos *casi* perfecta, es decir, teniendo todas las facultades necesarias para habérselas de manera eficaz con casi cualquier problema de manera semejante a como lo haría un ser humano de inteligencia media, que es lo que exige la noción de IA fuerte. En cuanto al modo de lograr semejante reproducción la unanimidad desaparece, y los investigadores se dividen en dos corrientes: los partidarios de la *IA subsimbólica*, que

defienden un enfoque materialista al estilo de Hobbes, y los partidarios de la *IA simbólica*, que defienden un enfoque dualista al estilo de Descartes, aunque sin reservar ninguna facultad como inimitable y exclusiva del ser humano. No es que estos últimos crean, en plena era espacial, que la realidad está dividida en dos sustancias, pero metodológicamente proceden como si tal división fuera cierta, pues suponen que es posible explicar la mente ignorando el cuerpo.

3. Computadoras electrónicas

Habíamos dejado la Historia de los autómatas del pensamiento en 1871, la fecha en que Charles Babbage murió sin haber podido completar la construcción de la máquina analítica, un invento que, por su diseño, habría sido el precursor de la computadora electrónica actual (Guijarro & González, 2010, p. 310). Pues bien, poco después, en 1889, el matemático norteamericano Herman Hollerith patentó la *Electric Tabulating Machine*, una máquina para el tratamiento de datos estadísticos que, al igual que la máquina analítica, utilizaba tarjetas perforadas como soporte de memoria. En sucesivos años, Hollerith fue añadiendo utilidades matemáticas a su invento hasta convertirlo en un artefacto de cálculo automatizado bastante versátil y eficaz. Con el objetivo de comercializarlo, en 1896 Hollerith fundó la Tabulating Machine Company, una empresa que en 1911 se fusionó con otras tres para crear la Computing Tabulating Recording Corporation (CTR). En 1924 la CTR cambió su nombre por el de International Business Machines Corporation, más conocida como IBM.

Las primeras máquinas de IBM eran más perfectas que cualquiera de las máquinas calculadoras anteriores. Ciertamente, no eran electrónicas todavía, ni tampoco de propósito general porque sus funciones estaban limitadas a cálculos matemáticos, pero sin embargo eran útiles para agilizar el trabajo de las redes de calculadores humanos, que todavía existían. Enseguida, los gobiernos más poderosos del mundo empezaron a utilizarlas para el tratamiento de grandes volúmenes de información. Un caso célebre lo encontramos en la Alemania nazi (Black, 2002). En 1933, tras haber alcanzado el poder, Hitler contrató los servicios de IBM para elaborar un censo de judíos y otros colectivos humanos que debían ser exterminados. El Tercer Reich tenía tantos enemigos, que la única forma de aniquilarlos era mecanizando todo

el proceso, tanto en el apartado material de la ejecución mediante las cámaras de gas, como en el intelectual mediante computadoras capaces de procesar los datos de millones de individuos. En agradecimiento por los servicios prestados, en 1937 el Führer impuso al entonces presidente de IBM, Thomas J. Watson, la Orden del Águila Alemana, una medalla que sólo se concedía a los más distinguidos simpatizantes del nazismo. Watson aceptó la condecoración, y no la rehusó hasta junio de 1940, cuando Hitler ya había invadido Polonia, Dinamarca, Noruega y, en la práctica, toda Francia mediante el Gobierno satélite de Vichy.

A pesar de sus virtudes, las máquinas de IBM seguían siendo artefactos mecánicos que realizaban las operaciones mediante el movimiento de partes físicas, mayormente relés, y eso era un inconveniente, pues la dependencia de un mecanismo físico disminuye la *fiabilidad*. Como ya hemos visto, el principal obstáculo técnico con el que tropezaron los diseñadores de autómatas matemáticos durante los siglos XVII y XVIII fue la precisión de los mecanismos para lograr máquinas infalibles, es decir, totalmente fiables. Un reloj puede atrasar unos segundos cada varias horas sin que ello suponga un problema grave, pero una máquina calculadora ha de arrojar siempre el resultado exacto. El reloj puede ser aproximativo; la calculadora ha de ser exacta. Sin embargo, el mundo físico no es exacto. Los dientes de un engranaje no son todos del mismo tamaño, y aun suponiendo que fueran producidos todos iguales utilizando tecnología de altísima precisión, el uso terminaría por desgastarlos de manera desigual. La única manera de evitar estos inconvenientes que reducen fatalmente la fiabilidad de una máquina computadora es construyéndola sobre un soporte físico que se comporte del modo más parecido posible al mundo de las Ideas de Platón, un lugar donde no exista el rozamiento, ni la fatiga de los materiales, ni los engranajes pueden atascarse. Ese soporte, en la actualidad, son los *circuitos electrónicos*, y su propiedad ideal deseada es la *digitalidad*.

Comúnmente se cree que la diferencia entre las máquinas mecánicas, como las de Babbage, y las electrónicas, como los ordenadores contemporáneos, es que las primeras son analógicas mientras que las segundas son digitales. Tal correspondencia

se puede conceder como verdadera siempre y cuando se tenga en cuenta que es sólo aproximativa. En rigor, la digitalidad es una propiedad exclusiva de los *sistemas formales*, y los sistemas formales no son reales, sino abstracciones intelectuales que representan determinados aspectos de la realidad al tiempo que ignoran otros. Para comprender por qué la digitalidad es una propiedad exclusiva de los sistemas formales es necesario definirlos detalladamente, diferenciándolos de los sistemas materiales. Vaya por delante que hay muchas definiciones posibles de "sistema formal" y "sistema material" (Ferrater, 1965a, p. 687), y que las que nosotros vamos a proponer no son, por tanto, las únicas, sino las ajustadas a los objetivos del análisis de las computadoras electrónicas, que es el tema de este capítulo.

Sistemas formales

José Ferrater Mora define un sistema como «un conjunto de elementos relacionados entre sí y armónicamente conjugados» (Ibíd., p. 687). A continuación distingue dos tipos de sistemas: *materiales* cuando los elementos son entidades reales, y *formales* cuando se trata de conceptos o enunciados. Hacia el final del artículo aborda la definición de sistema formal, diciendo que: «Un sistema formal es una serie de proposiciones dispuestas en tal forma, que de algunas de estas proposiciones, llamadas *axiomas*, se derivan otras proposiciones con ayuda de ciertas reglas de inferencia» (Ibíd., p. 689).

El filósofo norteamericano John Haugeland propone una definición de sistema formal que, aunque menos académica, es sin embargo equivalente a la de Ferrater Mora y mucho más accesible: «Un *sistema formal* es como un juego en el que las fichas (*token*, en el original en inglés, significa ficha, símbolo, señal) son manipuladas de acuerdo a reglas para ver qué configuraciones pueden ser obtenidas» (Haugeland, 1981a, p. 5). Haugeland propone el ajedrez como juego ilustrativo de su definición. Así, las proposiciones a las que se refiere Ferrater Mora, en la definición de Haugeland serían las posiciones de las piezas de ajedrez sobre el tablero, que pueden ser

codificadas en forma de proposiciones escritas en un notación como, por ejemplo, la algebraica, del tipo: Af1-c4 (alfil en la casilla f1 mueve a c4), Cb8-c6 (caballo en la casilla b8 mueve a c6), Dd1-h5 (dama en la casilla d1 mueve a h5), donde la letra mayúscula inicial indica el tipo de la pieza (P=Peón, T=Torre, C=Caballo, A=Alfil, D=Dama, R=Rey) y el par siguiente indica las coordenadas de la casilla, que se forman uniendo la letra de la columna (a-h) con el número de la fila (1-8). En cuanto a las reglas de inferencia a las que se refiere Ferrater Mora, serían las reglas de movimiento de las piezas. Aplicando las reglas a las proposiciones (o posiciones) iniciales, vamos avanzando en una partida de ajedrez, o bien vamos deduciendo proposiciones en un sistema formal como, por ejemplo, el del cálculo de predicados de primer orden: 1. $\neg(A \wedge B)$; 2. A; 3. $\neg B$. Las reglas del ajedrez permiten transformar Af1 en Ac4 (Af1-c4), pero no en Ae1 (Af1-e1), ya que eso implicaría cambiar de una casilla negra a una blanca, y no existe ninguna combinación de reglas que permita a los alfiles cambiar el color de las casillas que pisan.

A las definiciones de Ferrater Mora y Haugeland nosotros añadimos una precisión: que en los sistemas formales el conjunto de reglas de transformación ha de ser finito, y ellas han de ser perfectamente conocidas. De lo contrario, si hubiese infinitas reglas o se desconociese algunas de ellas, sería imposible jugar al ajedrez o calcular enunciados de lógica. En conclusión, podemos definir un sistema formal como un conjunto de elementos relacionados entre sí por un conjunto finito de reglas perfectamente conocidas. Así queda expuesto lo que es un sistema formal, pero todavía tenemos que distinguirlo del otro tipo de sistema: el material.

La máquina de diferencias de Babbage era un sistema formal en la mente de su creador: una serie de elementos relacionados entre sí de acuerdo a un conjunto finito de reglas perfectamente conocidas. Suponiendo que los planos eran correctos, en la mente de Babbage y de cualquiera que lo pensase el funcionamiento de aquel sistema formal era infalible. Sin embargo, al materializarlo el sistema formal pasó a convertirse en un *sistema material* o, más precisamente, la instanciación material de un sistema formal. Semejante salto del mundo de las Ideas al mundo real de las entidades físicas

dio lugar a que al conjunto de reglas establecido por Babbage se le unieran otras reglas no deseadas, pero inevitables, contra las que tuvo que luchar y que fueron la causa de que el proyecto se prolongase durante una década sin arrojar resultados satisfactorios. Esas reglas no deseadas eran las leyes de la naturaleza que afectan, quiérase o no, a la materia con la que él estaba construyendo su máquina, leyes tales como que el rozamiento erosiona las superficies, y que la transformación de la energía cinética en energía calorífica produce la dilatación de las piezas metálicas.

Aquí encontramos justamente la diferencia entre los sistemas formales y los materiales, de acuerdo a las definiciones que hemos formulado: el formal es una abstracción intelectual, mientras que el material es un sistema realmente existente en el mundo físico. Al ser una entidad del mundo físico, las reglas que rigen el funcionamiento de sus elementos pasan de ser finitas y perfectamente conocidas a ser, como mínimo, imperfectamente conocidas. En el mundo físico las reglas o leyes que actúan sobre la materia tal vez sean finitas, pero desde luego no son todas perfectamente conocidas, y si lo fueran, sería imposible saber que lo son. No hay manera de verificar un enunciado de una ciencia empírica, es decir, de determinar que es verdadero, pues siempre cabe la posibilidad de que se conozca una nueva experiencia que debería ser explicada por dicho enunciado y que, sin embargo, no la explique. Las ciencias empíricas no pueden determinar todas las reglas que afectan a la relación entre un conjunto de elementos.

Digitalidad

Por consiguiente, ningún sistema material es formal, y nada empírico es digital, dado que la digitalidad es una propiedad exclusiva de los sistemas formales. Sólo los sistemas formales son digitales gracias a su condición de sistemas inmateriales del mundo de las Ideas de Platón regidos por conjuntos finitos de reglas perfectamente conocidas. La física es un buen ejemplo de ciencia que opera constantemente con sistemas formales. Para calcular la trayectoria de un tiro parabólico se utilizan

ecuaciones, es decir, reglas que describen relaciones entre elementos como la masa, la velocidad y la inclinación. Pero esas ecuaciones no expresan todas las reglas que en la realidad están afectando a todos los elementos de un tiro parabólico. Por ejemplo, las ecuaciones no contemplan el color del proyectil, y sin embargo en el mundo real el color sí influye en la trayectoria, pues los colores oscuros como el negro absorben más fotones que los claros, y eso supone una diferencia de fuerza que en el mundo real está actuando y produce desviaciones, aunque sean de una magnitud del orden de ångströms (1 Å = $1x10^{-7}$ mm).

Nada de lo que existe en el mundo real es un sistema formal, ni siquiera los ordenadores. Ahora bien, aunque los ordenadores no son sistemas formales, sí son instanciaciones materiales de sistemas formales, gracias a lo cual poseen un cierto grado de digitalidad que varía en función del grado de semejanza de su materialidad con el mundo eidético en el que no existen cualidades físicas que puedan ser un obstáculo para el funcionamiento del sistema formal que instancian. En el caso de las computadoras electrónicas, dicho grado de semejanza es muy alto. Los flujos eléctricos que las hacen funcionar son, a la escala empleada actualmente en la fabricación de componentes informáticos, muy semejantes al mundo de las Ideas de Platón. Esto se aprecia con claridad mediante una comparativa. Supongamos tres computadoras: una construida con rollos de papel higiénico, la máquina de Babbage y un iPad. Parece el comienzo de un chiste, pero no lo es. Gracias al *principio de realizabilidad múltiple* (Copeland, 1993, p. 129), es posible construir con rollos de papel higiénico una computadora más potente que la suma de todos los iPad de última generación que hay en el mundo. No cabe duda de que sería un proyecto extravagante, pero posible al fin y al cabo. Comparativamente, el grado de digitalidad de la máquina de Babbage sería mayor que el del ordenador de rollos de papel higiénico, pues los engranajes de acero son menos susceptibles de fallar en su cometido que los rollos de cartón. Y, a su vez, el iPad sería más digital que la máquina de Babbage por la misma razón, y es que los flujos de electrones que lo hacen funcionar son más fiables que el movimiento de un sistema de piezas metálicas.

Sólo un sistema formal es digital en sentido estricto, y lo es porque cumple con los tres requisitos de la digitalidad: es autocontenido, es finitamente comprobable y está perfectamente definido (Haugeland, 1981a, p. 6). Que es *autocontenido* (*self-contained*) quiere decir que no le afecta lo que suceda en su entorno, es decir, que funciona de la misma manera haga frío, calor, o pintemos sus elementos del color que sea. Que es *perfectamente definido* (*perfectly definite*) significa que todos sus estados son exactos sin ambigüedad posible. En el ajedrez real cuando un jugador mueve una pieza, ésta pasa por varios lugares del espacio, incluyendo estados tan ambiguos como que la pieza pisa simultáneamente varias casillas. En cambio, en un sistema formal el movimiento es instantáneo desde la casilla de origen a la de destino, sin pasar por lugares intermedios. Al ser instanciaciones materiales de sistemas formales, las computadoras electrónicas también realizan sus operaciones a saltos. Por esta característica se dice que son *máquinas de estado discreto* (*discrete state machines*). Alan Turing las define como aquellas que «se mueven por saltos o golpes bruscos de un estado totalmente definido a otro. Estos estados difieren suficientemente para que podamos ignorar la posibilidad de confusión entre ellos. Hablando estrictamente estas máquinas no existen. En realidad todo se mueve de manera continua. Sin embargo, hay muchos tipos de máquinas de las cuales podemos pensar útilmente que son *máquinas de estado discreto*» (Turing, 1950, p. 18). Y, por último, que un sistema digital es *finitamente comprobable* (*finitely checkable*) quiere decir que se puede comprobar en un número de pasos finito que el estado actual es legal, lo que en el ajedrez se traduciría en que, por ejemplo, la legalidad de la posición de los alfiles de un mismo color se verifica comprobando que ocupan casillas de colores distintos.

Las primeras máquinas de IBM, las Hollerith, no eran digitales, como tampoco lo era la máquina de diferencias de Babbage, pues no eran autocontenidas en tanto que factores externos como la humedad afectaban a sus mecanismos, y tampoco eran perfectamente definidas, pues un relé, al ser un dispositivo electromecánico y no puramente electrónico, podía quedarse atascado en una posición similar a la de una pieza de ajedrez que pisa dos casillas simultáneamente. Por definición, es totalmente

imposible construir una calculadora infaliblemente precisa, y por tanto totalmente fiable, mediante un mecanismo analógico, es decir, un mecanismo que no sea digital. Como venimos recalcando, los ordenadores electrónicos tampoco son estrictamente digitales, pues no son sistemas formales, sino instanciaciones materiales de sistemas formales. Un cubo de agua es tan capaz de arruinar el funcionamiento del ordenador de rollos de papel higiénico como el del ordenador electrónico. Sin embargo, convencionalmente se concede que los ordenadores electrónicos son digitales debido a que, en condiciones normales, su grado de digitalidad es de los más elevados que se conocen en el mundo físico, gracias a que la electricidad *casi* funciona de un modo eidético, como las entidades del mundo de la Ideas.

Hay que reparar en que la digitalidad no es una prestación cualquiera de las computadoras electrónicas, sino la más importante. La digitalidad proporciona *fiabilidad*, en tanto que ningún factor empírico externo al objeto puede afectar a su perfecto funcionamiento, y la fiabilidad es la cualidad más importante de un utensilio. Como indica Martin Heidegger, la *fiabilidad* o el "ser de confianza" (*Verlässlichkeit*) es nada menos que la esencia del útil (Heidegger, 1952, p. 61). Analizando un cuadro de van Gogh que representa un par de zapatos viejos de campesino, Heidegger llega a la conclusión de que lo esencial de esos zapatos es que el que los usa pueda usarlos sin acordarse de ellos. Por tanto, lo esencial del útil es que sea fiable para que sea invisible. Sólo si es invisible podemos concentrarnos en la tarea que estamos realizando con el auxilio del útil. ¿Quién no ha experimentado alguna vez la irritación que produce el intentar escribir con un bolígrafo que escupe la tinta a tirones? Si te concentras en el bolígrafo, no puedes concentrarte en lo que quieres escribir con él. Es tan irritante como intentar hablar por teléfono con un contestador automático "inteligente": ambos utensilios fallan tanto que nos resulta imposible focalizar nuestra atención en la tarea. El útil ha de permanecer en el fondo para que el objetivo de nuestro interés pueda ocupar el primer plano. La necesidad de lo inconsciente como trasfondo de lo consciente impide traerlo todo a la conciencia, pensarlo todo al mismo tiempo como pretendía Descartes para evitar el error.

Este tercer capítulo vamos a dedicarlo a las computadoras electrónicas, soporte físico de la IA contemporánea. Dividiremos el tema en tres secciones, dedicadas respectivamente a las características formales, materiales y pedagógicas de estas máquinas. Cuando hablemos de computadoras electrónicas, o simplemente de computadoras u ordenadores, nos referiremos siempre, a menos que se indique otra cosa, a las de propósito general, que son las de máxima versatilidad, es decir, aquellas que por su diseño pueden simular cualquier computadora de propósito particular o computadoras *embebidas* (*embedded*) como, por ejemplo, una calculadora de bolsillo, una alarma antirrobo o el panel de control de un ascensor (Deitel & Deitel, 2012, p. 5).

En la primera sección expondremos las características formales de las computadoras electrónicas de propósito general mediante el examen de la máquina universal de Turing. Esto es posible porque, desde un punto de vista formal, una computadora electrónica de propósito general es equivalente a la máquina universal de Turing, en tanto que ambas son sistemas formales universales, y por tanto capaces de imitar de manera perfecta la ejecución de cualquier otro sistema formal (Haugeland, 1981a, p. 13), como por ejemplo el juego de ajedrez. En la segunda sección abordaremos las características materiales, que son las causadas por su naturaleza de objetos empíricos, sometidos a las limitaciones y contingencias del mundo real. Y, finalmente, en la tercera nos ocuparemos de las características pedagógicas, entendiendo por tales las destrezas cognitivas que sus características formales y materiales imponen a los seres humanos que las manipulan.

3.1. Características formales

En 1936, cuando los computadores todavía eran personas encerradas en cubículos, el matemático inglés Alan Turing publicó un artículo titulado *On computable numbers, with an application to the Entscheidungsproblem*. Algunos expertos opinan que éste es el texto fundacional de la ciencia moderna de la computación (Copeland, 2004, p. 6), mientras que otros atribuyen tal honor al informe *First draft of a report on*

the EDVAC elaborado en 1945 por el matemático estadounidense de origen húngaro John von Neumann (Ceruzzi, 1998, p.21). El propósito del artículo de Turing, como su título indica en inglés y en alemán, es doble: primero, definir formalmente la computación, y segundo, utilizar el resultado obtenido para resolver el problema de la decisión (*Entscheidungsproblem*). El medio ideado por el joven genio de tan sólo 24 años para acometer ambas cuestiones fue la descripción de una máquina hipotética que sería bautizada al año siguiente, 1937, como *máquina de Turing* por el también matemático Alonzo Church (Copeland, 2004, p. 6). Precisamente Church compitió con Turing por ser el primero en resolver el problema de la decisión. Al mismo tiempo que Turing, pero de manera independiente, Church desarrolló el *cálculo lambda* para llegar a la misma conclusión que Turing con su máquina: que el problema de la decisión es irresoluble. Ambos inventos, el cálculo lambda y la máquina de Turing, son el punto de despegue de la IA contemporánea, pues la máquina de Turing es la descripción formal de las computadoras electrónicas, mientras que el cálculo lambda sirvió al matemático John McCarthy del MIT como punto de partida para elaborar el lenguaje de programación LISP, el más utilizado, desde su publicación en 1958, por los pioneros de la IA para intentar crear las primeras computadoras inteligentes, hasta que cayó en desuso a finales de los 80, como veremos en el capítulo sexto.

Comenzaremos por describir la máquina de Turing, y a continuación veremos su aplicación al problema de la definición formal de la computación y al problema de la decisión, dos asuntos interesantes para el presente estudio en tanto que desvelan que las computadoras electrónicas tienen limitaciones formales, con lo que se desmiente la creencia popular de que son máquinas capaces de acometer cualquier tarea con la sola condición de que se les proporcionen los recursos materiales suficientes. Téngase en cuenta que la descripción de la máquina de Turing que vamos a realizar es una versión simplificada en la cual se han reemplazado algunos vocablos del artículo original de Turing por otros más actuales. Al tratarse de una máquina hipotética, perteneciente a un experimento mental, Turing no está obligado a explicar cuáles serían los mecanismos de ingeniería que la harían funcionar en la realidad.

La máquina de Turing

La máquina de Turing (T) es un sistema formal que se compone de tres partes: un cabezal, una cinta de longitud infinita (los objetos de longitud infinita son aceptables en los experimentos mentales) y una tabla finita de instrucciones. La *cinta* (*tape*) está dividida en casillas que pueden estar en blanco o llevar escrito un símbolo de un alfabeto finito que típicamente es el alfabeto binario de los símbolos 1 y 0. Así, cada casilla contiene un *bit*, acrónimo en inglés de *dígito binario* (*binary digit*). El *cabezal* (*scanner*) se desplaza por las casillas para leerlas de una en una, y tiene un dispositivo de registro que siempre se encuentra en un estado que puede ser cualquiera los posibles que figuran en la tercera y última parte de la máquina, la *tabla de instrucciones* (*table of instructions*). En cada momento, el cabezal puede realizar cualesquiera de las siguientes cuatro operaciones básicas o atómicas: borrar el contenido de la casilla actual, escribir un nuevo símbolo sobre ella, desplazarse hacia la casilla adyacente de la derecha o de la izquierda, y cambiar de estado. Las operaciones básicas a realizar por el cabezal en cada momento están determinadas por dos factores: el estado en el que se encuentre y el contenido de la casilla actual. Como se puede observar, la máquina de Turing es un sistema formal: los símbolos y el cabezal son los elementos relacionados entre sí de acuerdo a un conjunto finito de reglas perfectamente conocidas, que son las indicadas en la tabla de instrucciones. Las proposiciones iniciales son los símbolos y el estado del cabezal, que producen nuevas proposiciones mediante la aplicación de las instrucciones de la tabla.

Dejando al margen detalles menores, Turing completa la descripción de su máquina puntualizando que puede ser de dos tipos: *de elección* (*choice machines* ó *c-machines*) o *automática* (*automatic machines* ó *a-machines*). Las de elección se detienen en algún momento a la espera de que una fuente externa le suministre información para seguir computando, mientras que las automáticas son sistemas cerrados que trabajan sin interrupción hasta que terminan la computación y entonces

se detienen, en cuyo caso Turing las denomina *circulares* (*circular machines*), o bien no se detienen nunca, y reciben el nombre de *no-circulares* (*circle-free machines*). A Turing le interesan las automáticas no-circulares.

Una máquina como la que acabamos de describir (T) se dice que es una máquina *controlada por programa* (*program-controlled*), pues los pasos que efectúa están determinados por la lista de instrucciones, que es lo que hoy en día denominamos "programa". En noviembre de 1945 se finalizó la construcción de una de las primeras computadoras electrónicas de esta clase, el ENIAC, obra de un equipo de la Universidad de Pennsylvania de Estados Unidos dirigido por Presper Eckert y William Mauchly. La única utilidad del ENIAC era calcular trayectorias balísticas. Para que efectuara otra tarea era necesario modificar su lista de instrucciones, lo cual requería mover 6.000 interruptores. Un solo error en tan farragosa manipulación, y la máquina no funcionaría correctamente. Con el objetivo de solucionar este defecto, el siguiente paso en la evolución de las computadoras fue construir una de *programa almacenado* (*stored-program*), es decir, que tuviera el programa escrito no en la lista de instrucciones, sino en la propia cinta, de tal manera que un sencillo cambio de cinta permitiera cambiar de programa.

La máquina universal de Turing

La base teórica de tan gran avance fue proporcionada también por Turing en el mismo artículo de 1936, *On computable numbers* (Turing, 1936, p. 68). En aproximadamente treinta líneas, el genio inglés describió lo que hoy conocemos como la *máquina universal de Turing* (U), que es, como hemos señalado, nada más y nada menos que un sistema formal equivalente a cualquier computadora electrónica de propósito general (Ceruzzi, 1998, p. 149). En virtud de esta equivalencia, la máquina universal de Turing puede imitar a una computadora electrónica de propósito general, y a la inversa, una computadora electrónica de propósito general puede imitar a la máquina universal de Turing. De hecho, a los estudiantes de informática, y también a

los de psicología en aquellas universidades que todavía permanecen bajo el influjo del paradigma cognitivista del que hablaremos en el próximo capítulo, en algún momento temprano de su formación suele mandárseles la tarea de diseñar un programa de ordenador que imite el funcionamiento de la máquina universal de Turing, para que de esa manera comprendan qué es lo que en el fondo hacen las máquinas que tienen entre manos: hacen lo mismo que la máquina universal de Turing, con la única diferencia de que la máquina universal de Turing, gracias a su naturaleza teórica, tiene una capacidad de procesamiento superior a la de cualquier computadora electrónica, pues su cinta infinita le confiere una memoria ilimitada.

El primer ordenador electrónico construido según el diseño de programa almacenado de la máquina universal de Turing fue el Mark I, finalizado en la Universidad de Manchester de Inglaterra en junio de 1948, tres años antes de que los norteamericanos construyeran su propia versión, el UNIVAC, obra de los mismos ingenieros que crearon el ENIAC, Presper Eckert y William Mauchly. Recordemos que si Babbage hubiera finalizado su máquina analítica, aquélla habría sido la primera máquina universal de la Historia (Guijarro & González, 2010, p. 310). Ahora estamos en disposición de precisar que la diferencia entre la de Babbage y la universal de Turing estriba en que la de Babbage habría sido controlada por programa, pues el programa dictado por sus tarjetas perforadas no formaba parte del *store* en el que se almacenaban los datos computados, mientras que la universal de Turing es de programa almacenado (Copeland, 2004, p. 30).

El problema de la decisión

Como ya hemos dicho, el objetivo con el que Turing concibió su máquina, tanto la particular (T) como la universal (U), era doble: por una parte, definir formalmente la computación, y por otra, utilizar esa definición para resolver el problema de la decisión. Ambos asuntos están estrechamente vinculados. El *problema de la decisión* fue puesto de relieve por el matemático alemán David Hilbert a principios del siglo XX.

El objetivo de Hilbert era convertir las matemáticas en un sistema formal completo, consistente y decidible. *Completo* quiere decir que debe contener toda la verdad. *Consistente*, que debe contener sólo la verdad. Y *decidible*, que lo anterior debe poder comprobarse de forma algorítmica. Se dice que un conjunto es decidible cuando hay un algoritmo capaz de dirimir si una proposición cualquiera pertenece o no a dicho conjunto. La verdad, toda la verdad, y nada más que la verdad; un proyecto en busca de la verdad absoluta muy propio del carácter germánico.

Para disgusto de Hilbert y sus seguidores, en 1931 el checo Kurt Gödel demostró que el sistema formal de la aritmética no puede ser completo y consistente al mismo tiempo. En concreto, el primer teorema de la incompletud de Gödel establece que si el sistema formal de la aritmética es consistente, entonces no puede ser completo, es decir, que si sólo contiene proposiciones verdaderas, no puede contenerlas todas. En 1936 Turing y Church dieron la estocada definitiva al proyecto de Hilbert, demostrando la indecidibilidad de cualquier sistema formal de aritmética. Todo sistema formal de aritmética presupone un sistema lógico formal, por lo que demostrando la indecidibilidad de este último se demuestra la del primero. Eso es justo lo que hicieron Church y Turing: demostraron que el cálculo de predicados de primer orden, que es el sistema lógico formal más débil supuesto por cualquier sistema formal de aritmética, no es decidible (Ibíd., p. 49). Como ya hemos dicho, Church realizó su demostración inventando el cálculo lambda, mientras que Turing lo hizo con su máquina hipotética.

La demostración de Turing se divide en dos pasos: primero, formaliza la computación como procedimiento algorítmico, y a continuación utiliza esa definición formal para demostrar que no es posible decidir algorítmicamente si una proposición del cálculo de predicados de primer orden es un teorema. Un *teorema* es una *tautología*, una *verdad lógica*, una proposición verdadera en cualquier mundo posible. Un ejemplo de teorema es el principio de no contradicción, que establece que: $\neg(A \wedge \neg A)$, lo cual se lee como: no es verdad que algo sea verdadero y no verdadero (al mismo tiempo y en el mismo sentido, añadiría Aristóteles).

Hilbert pretendía encontrar un proceso para decidir *algorítmicamente* si una proposición de un sistema formal de aritmética es un teorema. Ya hemos visto anteriormente que en la Baja Edad Media la *algoritmia* era el nombre con el que se denominaba a las operaciones necesarias para la utilización de las nuevas cifras del sistema de notación indoarábigo. Los algoritmos eran, así pues, instrucciones. Con más precisión, en la actualidad un algoritmo se puede definir como un conjunto finito de instrucciones rutinarias cuya ejecución arroja un resultado deseado. *Rutinario* significa que el procedimiento es mecánico y por tanto es completamente objetivo, tan objetivo como el movimiento de la maquinaria de un reloj, cuyas instrucciones codificadas en forma de engranajes se ejecutan siguiendo una necesidad inexorable. Por tanto, lo que Hilbert deseaba era extirpar de las matemáticas al ser humano por ser una fuente de error, y así convertir la demostración de las verdades de las matemáticas en un proceso objetivo en tanto que rutinario.

La tesis Church-Turing

Sin embargo, el término "rutinario" es difícil de definir. Lo que para una persona puede ser rutinario, para otra puede suponer un esfuerzo de pensamiento creativo. Así definido, el concepto de algoritmo, piedra angular del proyecto hilbertiano para lograr la objetividad, depende en última instancia de una noción tan subjetiva como la rutina. Turing soluciona esta imprecisión con su máquina hipotética: la máquina de Turing es una definición formal de algoritmo sin implicaciones psicológicas (Ibíd., p. 43). ¿Qué es un algoritmo? Un algoritmo es aquel procedimiento que puede ser ejecutado por la máquina universal de Turing. Y si no puede ser ejecutado por la máquina universal de Turing, entonces no es un algoritmo, porque la máquina universal de Turing es capaz de ejecutar cualquier algoritmo.

Esta última sentencia es la *tesis Church-Turing*. Jack Copeland recoge dos de sus formulaciones más claras: «1. La máquina universal de Turing puede realizar cualquier cálculo realizable por un computador humano. 2. Cualquier método sistemático (es

decir, procedimiento efectivo o algoritmo) es efectuable por la máquina universal de Turing» (Ibíd., p. 41). No existe demostración matemática de la tesis Church-Turing, pero la comunidad científica está en su mayoría convencida de que es cierta: la máquina universal de Turing es capaz de efectuar cualquier algoritmo, y por tanto, una computadora electrónica provista de la cantidad de memoria suficiente es capaz de efectuar cualquier algoritmo. En favor de la tesis Church-Turing está el hecho de que lleva más de medio siglo resistiendo a la falsación, pues no se ha descubierto ningún algoritmo que no sea efectuable por una computadora electrónica. Cierto es que, desde el punto de vista de la lógica, la experiencia acumulada a favor de la tesis no permite concluir inductivamente que sea verdadera, pero desde un punto de vista psicológico su prolongada resistencia contribuye a su consolidación social. Las computadoras electrónicas llevan más de medio siglo ejecutando millones de algoritmos de manera perfecta, y eso fortalece la confianza, tanto dentro como fuera del ámbito científico, en que la tesis Church-Turing es verdadera.

Todos sabemos por experiencia propia que las computadoras electrónicas fallan, es decir, que a veces no hacen lo que deberían. No obstante, sus errores se deben siempre a causas totalmente ajenas a la tesis Church-Turing. El propio Turing distinguió los dos tipos de causas del malfuncionamiento de una computadora (Turing, 1950, p. 454). El primero es un error en el soporte material del sistema (*error of functioning*), como por ejemplo el sobrecalentamiento del procesador por un fallo del sistema de refrigeración. Y el segundo es un error cometido por un ser humano en la confección del algoritmo (*error of conclusion*), debido habitualmente a la imposibilidad de que una sola persona comprenda de manera simultánea, de un solo vistazo cartesiano sin participación de la memoria, la estructura completa de un algoritmo muy grande. Por ejemplo, el sistema operativo Windows Vista es un enorme algoritmo compuesto por 50 millones de líneas de código las cuales, obviamente, no fueron comprendidas en su totalidad por ninguno de los ingenieros que participaron en su elaboración. No pudo haber nadie que desempeñase el papel de "neurona obispo", omnisciente de la actividad de todos los demás.

Por tanto, si la tesis Church-Turing es verdadera, como parece que lo es, entonces la máquina universal de Turing, y por ende también su instanciación material en forma de computadora electrónica de propósito general, es capaz de realizar cualquier cálculo efectuable por un computador humano. Los computadores humanos no eran personas especialmente inteligentes, sino simples aplicadores de algoritmos, es decir, de instrucciones. Producían números igual que los operarios de una fábrica de agujas producen agujas, de manera rutinaria, mecánica. El perfil intelectual de los computadores de Maskelyne y De Prony era el de gente normal, desde amas de casa a cristaleros (Guijarro & González, 2010, p. 200). Sólo se exigía que tuvieran el hábito de hacer cálculos, pues el hábito automatiza los procesos mentales y por tanto disminuye la probabilidad de cometer un error.

La genialidad de la máquina de Turing a este respecto reside en que es la reducción a un sistema formal de *cualquier* tarea de computación realizable por los computadores humanos. Además, las operaciones básicas o atómicas de dicho sistema formal son sólo cuatro, y son totalmente mecánicas: borrar el contenido de la casilla actual, escribir un nuevo símbolo sobre ella, desplazarse hacia la casilla adyacente de la derecha o de la izquierda, y cambiar de estado. Turing se refiere constantemente a los computadores humanos para equiparar sus capacidades con las de su máquina universal (Turing, 1936, p. 76). Su propósito es formular una definición completamente formal del concepto de algoritmo, sin apelar a una noción psicológica tan imprecisa como la "rutina".

La sospecha de circularidad puede aparecer en este punto, pues si definimos los algoritmos como aquello que las máquinas de Turing hacen, entonces resulta obvio que las máquinas de Turing son la definición operacional de los algoritmos. Pero la ilusión de circularidad se rompe cuando reparamos en que Turing no pretende construir el concepto de algoritmo en base a su máquina, sino su máquina en base a lo que de hecho hacen las personas que realizan algoritmos, esto es, los computadores humanos. Si se encontrara un algoritmo que fuera ejecutable por un computador humano pero no por la máquina universal de Turing, entonces la tesis Church-Turing

quedaría refutada. Por tanto, la tesis Church-Turing no es circular, no incurre en petición de principio, sino que cumple con el criterio científico de la falsabilidad (Popper, 1934, p. 40), según el cual para que una proposición empírica sea científica ha de poder demostrarse que es falsa.

La máquina universal de Turing hace lo que haría un computador humano, y al mismo tiempo es un sistema formal capaz de imitar el funcionamiento de cualquier computadora electrónica. Estas equivalencias se deben a que los computadores humanos, las computadoras electrónicas y la máquina universal de Turing son tres sistemas que sólo hacen una cosa: efectuar algoritmos. En el siglo XIX, De Prony y Maskelyne concibieron los algoritmos como herramientas para producir solamente números. Hoy en día sabemos que los algoritmos sirven para hacer muchas otras cosas. De hecho, todo lo que hacen las computadoras electrónicas es producto de algoritmos: desde la simulación de instrumentos musicales hasta las imágenes por ordenador indistinguibles de la realidad. Después de todo, resulta que, analizada con detenimiento, la creencia infantil de que hay unos señores diminutos en el interior del ordenador que obedecen lo que se les manda es, desde un doble punto de vista histórico y metafórico, totalmente verdadera.

En teoría, podríamos crear una película de animación digital empleando una red de computadores humanos en vez de computadoras electrónicas. Sólo habría que suministrarles las órdenes adecuadas a través de un terminal que hiciera las veces de pantalla y teclado. Sería un proyecto tan loco como el de la computadora construida con rollos de papel higiénico, pero sería posible, y pensar en dicha posibilidad es una manera de recordar que las computadoras electrónicas no son artefactos mágicos capaces de hacer cualquier cosa en virtud de las misteriosas propiedades de la electricidad que fluye por sus circuitos (Turing, 1950, p. 446). La realidad es que sólo son máquinas que sustituyen a los computadores humanos en la única tarea que éstos debían hacer: efectuar algoritmos. Ciertamente, las computadoras electrónicas han demostrado que los algoritmos son capaces de producir una gran variedad de resultados, pero eso no demuestra que puedan arrojar cualquier resultado.

Limitaciones de las máquinas de Turing

Que la máquina universal de Turing sea capaz de efectuar cualquier algoritmo y que los algoritmos por definición arrojen resultados deseados, no son premisas de las que se siga que la máquina universal de Turing es capaz de arrojar cualquier resultado deseado, pues en la lógica y las matemáticas hay resultados que no pueden ser producidos algorítmicamente. Por ejemplo, la máquina universal de Turing no puede decidir si cualquier proposición de la lógica de predicados de primer orden es un teorema, tal y como se demuestra en la última sección de *On computable numbers*, pues se trata de un sistema lógico cuyas proposiciones forman un conjunto *semidecidible*, es decir, que existe un algoritmo tal que, dada una proposición que es un teorema, es capaz de determinar que, en efecto, se trata de un teorema, pero dada una proposición que no lo es, puede ocurrir que, o bien el algoritmo se detenga en algún momento para indicar que no lo es, o bien siga funcionando durante un tiempo indefinido, con la incertidumbre de no saber si se detendrá o no, y por tanto sin saber si la proposición es o no un teorema.

Las máquinas de Turing, al igual que las computadoras electrónicas, producen resultados numéricos que en símbolos binarios son secuencias de 1 y 0. Esas secuencias son finitas en el caso de las máquinas circulares, e infinitas en el de las no-circulares. Un ejemplo de estas últimas sería una máquina que calculase los decimales de π, un proceso sin final. De acuerdo a la peculiar noción de computabilidad que maneja Turing, sólo son computables las secuencias producidas por máquinas no-circulares. Para demostrar que existen secuencias infinitas que no son computables, el genio inglés pide que imaginemos una tabla en la que figuran todas las secuencias computables. Cada fila de la tabla es una de las secuencias computables, y cada columna muestra las cifras ordenadas de dichas secuencias. El número de filas es infinito, pues hay infinitas secuencias computables, y el número de columnas también es infinito, pues las secuencias computables no tienen final, como la del número π. A

continuación, dice Turing, tomemos la primera cifra de la primera fila, la segunda cifra de la segunda fila, y así sucesivamente. El resultado sería la secuencia diagonal de la tabla, a la que Turing da el nombre de secuencia β', y sería del estilo: 100110100... El último paso consiste en hallar β, que es el complemento de β'. La secuencia β se obtiene cambiando los 1 por 0 y los 0 por 1 de la secuencia β'. Por tanto, la secuencia complementaria β sería: 011001011... β no está en la lista de las secuencias computables, pues difiere de todas las secuencias computables en al menos una cifra. Por tanto, es una secuencia infinita no computable, o lo que es igual: no producible algorítmicamente.

Con este *argumento diagonal* y, en aplicación de la tesis Church-Turing, se colige que β no puede ser producida por la máquina universal de Turing ni por ninguna computadora electrónica. Hay quien podría objetar que β sí es computable, en tanto que ha sido obtenida mediante el algoritmo recién descrito. Pero Turing refuta esa objeción. Llamemos BETA a la máquina universal destinada a tal propósito imposible (Copeland, 2004, p. 37). BETA tendría que simular a todas las máquinas de Turing que producen secuencias computables para producir dichas secuencias hasta llegar al dígito deseado de cada una: el primero de la primera, el segundo de la segunda, etc. Sin embargo, en tal proceso llegaría un momento en el que BETA tendría que simularse a sí misma, pues se supone, según la objeción, que ella produce una secuencia computable. Al ocurrir esto, BETA tendría que simularse a sí misma simulando a todas las máquinas que producen las secuencias anteriores, hasta volver a llegar al punto en que tendría que simularse a sí misma de nuevo. Y así hasta el infinito. En conclusión, la secuencia β no es computable y BETA es imposible. La producción de β es descriptible mediante un procedimiento preciso que parece un algoritmo pero que, en realidad, no lo es, porque no conduce al resultado deseado.

Así queda probado que hay resultados que no pueden ser obtenidos mediante algoritmos, y por consiguiente no pueden ser producidos por ninguna computadora electrónica. Esta limitación formal quizás suponga un obstáculo insalvable para la IA, y por tanto podría ser la explicación teórica de la *imperfectibilidad insalvable*

proclamada por los mitos (Robinet, 1973, p. 38). Es una cuestión que abordaremos en el capítulo séptimo. Sin ir tan lejos de momento, el objetivo que habíamos propuesto para esta sección era examinar las características formales de las computadoras electrónicas. La abrumadora eficacia exhibida por estas máquinas ha contribuido a extender la creencia de que no hay nada que se les resista, que cualquier dificultad pueden superarla con procesadores más rápidos y memorias más grandes. Sin embargo, hemos visto que se trata de una creencia falsa. La máquina universal de Turing, por su condición de experimento mental, es tan rápida como se desee y tiene memoria infinita, pero incluso así hay cosas que no puede hacer. Las instanciaciones materiales equivalentes a ella en forma de computadoras electrónicas están sometidas, por tanto, a las mismas limitaciones. Una vez expuestas las características formales de las computadoras electrónicas, pasemos a las materiales.

3.2. Características materiales

Las características materiales de las computadoras electrónicas son las causadas por su condición de objetos del mundo real. La parte material de una computadora electrónica se denomina *hardware* (literalmente, *conjunto de utensilios duros*) mientras que los algoritmos que ejecuta se denominan *programas* o *software* (*conjunto de utensilios blandos*). En la actualidad, el principal componente del hardware son los circuitos integrados, cuyo funcionamiento a escala electrónica evita una gran cantidad de problemas técnicos encontrados en el pasado, como por ejemplo el de la transmisión de decenas, que se tardó más de un siglo en resolver desde que Shickard tropezó con él en 1623 hasta que Hahn lo solucionó en 1785. Pero no los evita todos, razón por la cual hemos señalado que ninguna computadora, ni siquiera las electrónicas, es digital en sentido estricto. En 1948 Turing calculó que, suponiendo que el hardware no se estropease nunca, el error de lectura de un *bit* en las computadoras diseñadas con la tecnología de su época era de 1 entre 10 elevado a 100.000 billones, a causa de un fenómeno físico denominado *agitación termal del voltaje* (Turing, 1948, p.

414). En nuestros días Western Digital, uno de los fabricantes más prestigiosos de discos duros, garantiza que sus soportes de almacenamiento de datos VelociRaptor de 2,5'' alcanzan los 1,4 millones de horas de funcionamiento sin fallos. Es mucho, sin duda, pero la falibilidad está ahí y es inevitable.

El hardware es *falible*, y cuando falla, la consecuencia habitual es la imposibilidad de ejecutar los algoritmos codificados en el software. Puede romperse una soldadura, quemarse una resistencia, o sobrecargarse la fuente de alimentación a causa de un pico de voltaje. Por tanto, el funcionamiento electrónico ha aumentado la digitalidad respecto de máquinas más antiguas como las de Shickard y Babbage, o lo que es lo mismo, ha disminuido la exposición a las contingencias empíricas, pero no la ha eliminado por completo. Esto supone un gran problema para la IA, pues, dada la arquitectura de las computadoras electrónicas actuales, la más pequeña avería en el hardware suele resultar en un error fatal, mientras que en el cerebro mueren miles de neuronas cada día sin que por ello perdamos recuerdos o destrezas intelectuales. Aunque, según se mire, el hecho de que el error sea casi siempre fatal puede ser considerado como un mecanismo de seguridad, porque garantiza que una avería no producirá un error persistente y silencioso en, por ejemplo, la transmisión de decimales, sino que el sistema entero dejará de funcionar. Es preferible un ordenador que o funcione bien o no funcione, antes que un ordenador al que las averías lo afecten degradando su buen funcionamiento: hoy comete un error, mañana dos, y cuando creemos que el misil está cerca del objetivo, resulta que no impacta en el Kremlin, sino en el aparcamiento de enfrente.

Además de ser falible, el hardware es *limitado*. La máquina de Turing, gracias a su condición de experimento mental, tiene dos características imposibles de materializar: velocidad y memoria ilimitadas. Puede funcionar a una velocidad arbitrariamente alta y dispone de una cinta de memoria de longitud infinita. Ambas características garantizan que, si la tesis Church-Turing es verdadera, la máquina universal de Turing puede efectuar cualquier algoritmo. En cambio, las computadoras electrónicas operan a velocidades limitadas y con memorias limitadas. Una de las

excusas que tradicionalmente han presentado los investigadores de la IA para justificar su falta de progresos en la construcción de inteligencias artificiales en sentido fuerte ha sido que la tecnología de la época no era lo suficientemente rápida y que no ofrecía la suficiente capacidad de almacenamiento para ejecutar los algoritmos responsables de producir la inteligencia humana (Hawkins & Blakeslee, 2004, p. 83). Semejante queja era razonable en los años 50, cuando los pioneros como Allen Newell intentaron crear las primeras inteligencias artificiales, pero no en la actualidad. Desde el primer ENIAC, de 1945, las computadoras han aumentado vertiginosamente sus prestaciones gracias a dos técnicas de ingeniería: la miniaturización y la suma de componentes.

Miniaturización

Miniaturizar es el proceso tecnológico para reducir el tamaño de los dispositivos electrónicos. El ENIAC, la computadora más avanzada de su época, pesaba 27 toneladas y ocupaba más de 150 metros cuadrados. En cambio, un iPad de 2012 pesa 700 gramos, cabe en la palma de la mano y tiene una potencia muy superior. El enorme tamaño del ENIAC se debía a que estaba construido con aparatosos tubos de vacío, el componente principal de la primera de las varias generaciones en las que habitualmente se divide la evolución de las computadoras electrónicas. Poco después, los tubos de vacío fueron sustituidos por los transistores, mucho más pequeños, y a finales de los 60 llegó la tercera generación, la de las computadoras basadas en circuitos electrónicos integrados, también llamados *chips* o *microchips*, unos dispositivos aún más pequeños que son los que usamos en la actualidad. La cuarta generación abarcó la década de los 70, siendo su rasgo distintivo la comercialización de los primeros microprocesadores, que son un tipo de circuito integrado. Durante los 80 se desarrolló la quinta generación, la de los ordenadores personales o *PC* (*personal computer*). Y, finalmente, en los 90 comenzó la sexta, que es en la que todavía nos hallamos inmersos. Desde cierto punto de vista, las generaciones cuarta, quinta y sexta pueden ser clasificadas como etapas menores dentro de la tercera, en tanto que han

seguido basándose en la tecnología de circuitos integrados (Ceruzzi, 1998, p. 6). La razón por la cual merecen ser nombradas como generaciones con entidad propia es porque han contribuido a disminuir el tamaño y el precio mediante avances en las técnicas de miniaturización y de producción en serie.

La *miniaturización* y el *abaratamiento* de las computadoras electrónicas han sido los factores materiales más importantes para su inserción social. Recordemos, por ejemplo, que la máquina de Hahn no tuvo éxito, entre otras razones, por culpa de sus elevados costes de producción. Si las computadoras electrónicas no hubieran reducido su tamaño y su precio tan pronunciadamente como lo han hecho, entonces los ordenadores personales no existirían, y las computadoras seguirían siendo enormes *máquinas centrales* (*mainframes*), como UNIVAC, ocultas en los sótanos de instalaciones gubernamentales. Gordon Moore, cofundador de Intel, la empresa que inició la cuarta generación con el lanzamiento del microprocesador 4004, enunció en 1965 su famosa *ley de Moore*, la cual pronosticaba que aproximadamente cada año y medio se duplicaría la potencia de los circuitos integrados (Deitel & Deitel, 2012, p. 6), o dicho a la inversa, que un circuito integrado de la misma potencia reduciría su tamaño a la mitad. Moore acertó, y su ley lleva cumpliéndose desde entonces. Paralelamente, la miniaturización ha venido acompañada de un descenso vertiginoso de los costes de fabricación. El primer ordenador electrónico de IBM, el modelo 701, se empezó a comercializar en 1953 a un precio de un millón de dólares la unidad, lo mismo que costaba su competidor, el UNIVAC. Era tanto dinero para la época que, en lugar de venderlo, IBM lo alquilaba a sus clientes (Ceruzzi, 1998, p. 35).

Las computadoras electrónicas son cada vez más potentes, pequeñas y baratas. Desde los años 60, el precio por ciclo de computación ha caído un 99,9% (Carr, 2010, p. 83). Es una tendencia que a ojos del consumidor parece ilimitada, pero en realidad su final está muy cerca. Hoy en día los circuitos integrados para computadoras comerciales se fabrican con una resolución aproximada de 30 nanómetros (1 nm = $1x10^{-6}$ mm), lo que significa que la unidad de medida empleada por los ingenieros para trazar los canales por los que circula la electricidad es de esa magnitud. Por ejemplo,

los microprocesadores más modernos de Intel a mediados de 2012, los Ivy Bridge, estaban fabricados en 22 nm. Los físicos dicen que el límite está en 10 nm (Kaku, 2011, p. 70). Semejante barrera es, por el momento, infranqueable debido a la longitud de onda de la luz ultravioleta, que es el instrumento utilizado para grabar el dibujo de los circuitos sobre las placas de silicio. Y si se llegara a traspasar, entonces los ingenieros tendrían que enfrentarse a los problemas derivados de operar a escala atómica. El más importante de ellos es el *principio de incertidumbre de Heisenberg*, el cual establece que es imposible conocer a la vez la posición y la velocidad de una partícula. En ese escenario, los electrones que circulasen por los circuitos se escaparían de las delgadas líneas atómicas de silicio trazadas para su flujo ordenado, provocando cortocircuitos.

Suma de componentes

La otra estrategia para aumentar la potencia de las computadoras electrónicas es la suma de componentes. En este punto tenemos que distinguir los dos tipos de arquitecturas empleadas en la construcción de computadoras en la actualidad: la *von Neumann* y las *no-von*. La arquitectura von Neumann fue propuesta por el matemático John von Neumann en un informe de 1945 que ya hemos citado antes: *First draft of a report on the EDVAC*. Las características más importantes de la arquitectura von Neumann son tres (Ceruzzi, 1998, p. 23). En primer lugar, la unidad que procesa la información está separada de la unidad que la almacena, una idea que ya se le había ocurrido a Charles Babbage un siglo antes cuando decidió dividir su máquina analítica en los módulos *mill* (procesador) y *store* (almacén). En segundo lugar, la tabla de instrucciones y los datos de procesamiento han de estar almacenados en el mismo dispositivo, una idea que tampoco es original de von Neumann, pues es justamente la característica distintiva de la máquina universal de Turing, descrita en 1936 en el artículo *On computable numbers*. Y, en tercer lugar, el procesador ejecuta las instrucciones *en serie* (*serial processing*), es decir, una detrás de otra, y preferiblemente sin que el cabezal lector tenga que dar saltos a posiciones de memoria

distintas de la contigua a la actual (Ibíd., p. 42). Como se puede observar, von Neumann no inventó todas las características de la arquitectura que lleva su nombre, pero sí tuvo la audacia de seleccionar lo mejor de los mejores para reunirlo en una sola gran idea (Copeland, 2004, p. 22). Como decía Woody Allen citando supuestamente a Picasso: «Los artistas buenos copian; los grandes, roban». Como quiera que lo hiciese, el nombre del gran von Neumann designa una arquitectura que fue la dominante durante medio siglo en las computadoras electrónicas, desde sus inicios en 1945 hasta 1995 (Ceruzzi, 1998, p. 21).

A partir de 1995 comenzaron a surgir comercialmente las arquitecturas no-von, y con ellas empezó la actual generación de computadoras, la sexta. Las arquitecturas no-von pueden ser de muchos tipos: de hecho, por definición, toda arquitectura que no sea von Neumann es no-von. Un tipo de arquitectura no-von muy prometedor para el futuro es la de las computadoras basadas en las leyes de la física cuántica, que de momento están en fase experimental. No obstante, a día de hoy las únicas no-von operativas y fiables son las computadoras *multinúcleo*, que disponen no de una, sino de varias unidades de procesamiento de información. Cada núcleo procesa una parte del programa almacenado *en paralelo* (*parallel processing*) a los demás, con lo cual, a más núcleos, mayor velocidad. A pesar de esta diferencia arquitectónica, las von Neumann y las no-von multinúcleo son formalmente equivalentes, dado que lo único que hacen ambas es efectuar algoritmos, aunque lo hagan de manera distinta. Las von Neumann son equivalentes a la máquina universal de Turing, gracias a lo cual, y según la tesis Church-Turing, no hay algoritmo que no puedan efectuar si se les proporciona la memoria y el tiempo necesarios. Por tanto, cualquier algoritmo efectuable en paralelo por las no-von multinúcleo es también efectuable en serie por las von Neumann, con la única diferencia de que éstas últimas tardarán más.

La arquitectura no-von de múltiples núcleos y múltiples soportes de memoria es precisamente la estrategia empleada por los ingenieros para construir los superordenadores. En 1997, Deep Blue era el superordenador más potente del mundo, gracias a sus 30 nodos con 30 procesadores en cada uno. A fecha de marzo de 2012, el

superordenador más potente era el Tianhe-1A, del Centro Nacional de Supercomputación de China, con más de 180.000 núcleos de procesamiento, según el portal top500.org. Comparativamente, en sólo 15 años el número de núcleos del mayor superordenador del mundo se ha multiplicado por 6.000. Sin embargo, Tianhe-1A sigue siendo sólo un gran sistema experto que está tan lejos de la inteligencia humana como lo estaba Deep Blue. Deep Blue jugaba al ajedrez, y Tianhe-1A, como mucho, simula fenómenos climáticos, pero sigue sin ser capaz de mantener una conversación coherente, ni siquiera sobre algo tan banal como su especialidad: charlar sobre lo mucho o lo poco que llueve últimamente.

En definitiva, gracias a la miniaturización y a la suma de componentes, la ingeniería informática ha alcanzado en pocas décadas un grado de desarrollo muy elevado. Es la tecnología que, en toda Historia de la humanidad, más ha progresado en menos tiempo, tanto a nivel técnico como de implantación social. En la actualidad disponemos de computadoras electrónicas muy potentes, pero no parece que un futuro próximo vayan a mejorar aún más sus prestaciones, pues la ley de Moore está a punto de extinguirse. Por tanto, ésta es la tecnología con la que los investigadores de la IA deben construir inteligencias artificiales en sentido fuerte. Y a la inversa: para producir inteligencias artificiales en sentido fuerte, éstas son las computadoras que deben ser manipuladas por los seres humanos. Así enlazamos con las características pedagógicas de las computadoras electrónicas, es decir, las destrezas cognitivas que estas máquinas imponen para ser manipuladas.

3.3. Características pedagógicas

Para comprender las características pedagógicas de las computadoras electrónicas debemos comenzar examinando estas máquinas en su condición de instrumentos producidos por la técnica. La *técnica*, del griego *tékne* (τέχνη), puede definirse como «toda serie de reglas por medio de las cuales se consigue algo» (Ferrater, 1965b, p. 763). A la luz de esta definición se observa que la técnica está

estrechamente relacionada con el concepto de razón instrumental visto antes. La razón instrumental, como dijimos en el primer capítulo, es aquella que se ocupa de maximizar el rendimiento de los medios para alcanzar un fin, sin cuestionar la racionalidad del fin en sí mismo. Por tanto, la técnica es producto de la razón instrumental. Y, a su vez, las herramientas son productos de la técnica. Habitualmente se confunde la técnica con la tecnología y con las herramientas (Spengler, 1931, p. 14). Esta confusión es incorrecta porque, en primer lugar, una tecnología puede ser un conjunto tanto de técnicas como de herramientas. Y, en segundo, la técnica no debe identificarse con las herramientas. La técnica, como señala Oswald Spengler, no se trata tanto de la fabricación de cosas como del manejo de ellas (Ibíd., p. 14). En cualquier caso, todas las técnicas y tecnologías, sean con o sin herramienta, se dividen en dos categorías: físicas y sociales. Nótese que esta distinción corresponde a la que ya vimos en el primer capítulo cuando, al hablar de las competencias de los sistemas expertos, dividimos el mundo en dos dimensiones: física y social.

La técnica no es exclusivamente humana, pues los animales también la tienen. Por ejemplo, ciertas aves demuestran una técnica física portentosa en la fabricación de nidos, y los chimpancés exhiben técnicas sociales para manipular la voluntad de sus semejantes y conseguir de ellos lo que desean. Descartes observó este hecho para fundamentar su convencimiento de que las inteligencias de dominio específico sí pueden ser replicadas perfectamente por los autómatas, aunque reservando la inteligencia de alcance general para el hombre.

Podemos señalar al menos dos diferencias esenciales entre la técnica humana y la animal. La primera es que los animales sólo tienen *técnica de la especie*, heredada tras un largo proceso de selección natural (Ibíd., p. 31). Los rasgos conductuales de un individuo, como por ejemplo la técnica de construcción de nidos, forman parte de su fenotipo. El *fenotipo* es resultado de la interacción de factores genéticos y ambientales, y a su vez, los factores genéticos están determinados por los genes de los progenitores y por mutaciones espontáneas. Si un cierto fenotipo es eficaz para la supervivencia, entonces los genes que están en la base de su expresión se propagan

más que los de otros individuos, hasta el punto de que, en ocasiones, dichos rasgos fenotípicos llegan a ser propios de toda la especie, como es el caso de la construcción de nidos en las cigüeñas. No existen cigüeñas que no sepan construir nidos, porque las que alguna vez hubo se extinguieron en la competición con las que sí habían heredado esa técnica. Los seres humanos, en tanto que animales, también heredamos mediante los genes una técnica de la especie, pero además tenemos una *técnica individual* que es inventiva, aprendible y susceptible de desarrollo (Spengler, 1931, p. 31). La técnica individual la aprendemos de las generaciones anteriores por imitación y a través del lenguaje, el cual, según la teoría de la gramática universal de Chomsky y la teoría modular de la mente (García, 2001a, p. 200), es una predisposición de la especie que se ha consolidado en nuestro genotipo gracias a su eficacia para la supervivencia. Debido al efecto Baldwin, del que hablaremos en el capítulo sexto, tenemos una habilidad innata para el lenguaje, igual que las cigüeñas la tienen para construir nidos.

La segunda diferencia esencial entre la técnica humana y la animal es que la nuestra produce herramientas y las desarrolla. Cierto es que algunos animales ocasionalmente manipulan objetos para convertirlos en herramientas con las que realizar determinadas acciones. Por ejemplo, los chimpancés utilizan palos largos para extraer la miel de los panales. Pero esa habilidad no es producto de una inteligencia específica, como sí lo es en la especie humana, sino de una inteligencia general, lo cual implica limitaciones insalvables para nuestros ancestros. La inteligencia, de acuerdo a sus dominios de aplicación, se divide en *inteligencia general* y un conjunto de *inteligencias específicas*. Las específicas son aquellas que han sido modeladas a través del curso de la evolución para hacer frente a problemas habituales cuya resolución supone una ventaja para la supervivencia, mientras que la inteligencia general es la que se ocupa de resolver el resto de problemas. Las inteligencias específicas pueden ser *físicas* o *sociales*. Algunos autores se refieren a las inteligencias físicas como inteligencias técnicas (Ibíd., p. 180), pero no nos parece adecuado denominarlas así, en tanto que induce a pensar que la técnica es exclusiva del ámbito físico, cuando en realidad también existen técnicas sociales.

Está demostrado que los chimpancés tienen inteligencia específica social, concretada en sus correspondientes técnicas, dado que la inteligencia social es la clave en el proceso de hominización y desarrollo de la mente (Ibíd., 175). Sin embargo, parece ser que no poseen una inteligencia específica física. En consecuencia, los problemas del mundo físico los resuelven aplicando una inteligencia general que no ha sido modelada evolutivamente para ese ámbito en particular y que, por tanto, tiene un rendimiento menor. «Cuando un chimpancé ve que otro introduce un palo en el hormiguero y al sacarlo se come las termitas adheridas, luego empieza a hacer algo parecido hasta lograrlo después de varios ensayos. No parece que sea propiamente un aprendizaje por imitación, si por tal entendemos que ha captado la tarea, el objetivo de la acción y los medios para lograrlo. Por eso en más de treinta años de observación de tales conductas en los chimpancés no se ha constatado avance alguno: cada generación de chimpancés se esfuerza para alcanzar tan sólo el nivel técnico (en el ámbito físico) de la generación precedente» (Ibíd., p. 181). En cambio, los seres humanos creamos herramientas y las desarrollamos de manera acumulativa a nivel de la especie, una generación tras otra.

Las tecnologías son, como hemos indicado, medios para obtener fines. En función de la naturaleza de dichos fines las tecnologías pueden dividirse en cuatro tipos (Carr, 2010, p. 44). El primero comprende aquellas que ayudan a ampliar la fuerza, el campo de acción o la precisión del *sistema motor*. En este conjunto podemos enumerar inventos como la máquina de vapor, las armas de fuego y los brazos robóticos que fabrican circuitos integrados a escalas nanométricas. El segundo es el de aquellas que aumentan la agudeza o el alcance del *sistema sensorial*. En este grupo se inscriben inventos como las gafas, el microscopio y el radar. El tercer tipo es el de aquellas que satisfacen nuestros *deseos* mediante la modificación de la naturaleza. Algunos de sus ejemplos serían la píldora anticonceptiva y el maíz transgénico. Y, finalmente, el cuarto grupo es el de las tecnologías intelectuales, que son aquellas que auxilian o potencian las *facultades mentales*. En este apartado podemos enumerar el lenguaje, la escritura, la imprenta, los ábacos y las computadoras electrónicas, capaces

éstas últimas de tareas tan diversas como extraer conclusiones estadísticas a partir de enormes volúmenes de datos o realizar predicciones meteorológicas operando sobre miles de variables en ecuaciones matemáticas tan complejas que serían imposibles de resolver por los seres humanos en un tiempo útil.

Lo que la técnica impone

Los seres humanos inventamos tecnologías con el objetivo de que nos sirvan para alcanzar un fin. Este carácter de medio ciego respecto de los fines da lugar a la creencia errónea pero muy extendida de que las tecnologías son neutrales, es decir, ni buenas ni malas, sino que su calificación moral depende de los fines para los que se utilicen. Según esa creencia, las computadoras electrónicas serían buenas cuando se emplean para fines buenos como, por ejemplo, comunicarnos con nuestros amigos a distancia, y en cambio malas cuando se utilizan, pongamos por caso, para exterminar a colectivos humanos, como ya hemos mencionado que hizo Hitler con las máquinas computadoras de IBM (Black, 2002).

Lo cierto es que ninguna técnica o herramienta es neutral, sino que toda tecnología es *pedagógica*, educativa en sentido amplio (Weizenbaum, 1976, p. 26). Pensemos en una herramienta rudimentaria como un martillo. En el momento en que lo cogemos con la mano, el martillo es literalmente incorporado al mapa neuronal de nuestro cuerpo (Hawkins & Blakeslee, 2004, p. 76). Cuanto más lo utilicemos, más perfecto será dicho mapa y con más soltura manejaremos el martillo. Es el mismo proceso por el cual un niño va adquiriendo destreza en el manejo de su propio cuerpo, y que continúa a lo largo de la vida adulta a medida que ampliamos nuestro cuerpo y nuestra mente con tecnologías de los cuatro tipos descritos. Si tenemos el hábito de conducir un coche pequeño, cuando cambiamos a otro de mayores dimensiones aumenta la probabilidad de que rocemos la carrocería. El motivo es que nuestro cerebro todavía está operando con el cálculo espacial del coche pequeño. Un caso aún más extremo que los del martillo y el coche es el del ojo supernumerario añadido

mediante técnica quirúrgica (Kandel, Schwartz & Jessell, 1995, p. 478). La neurocientífica Martha Constantine-Paton y su equipo implantaron a una rana recién nacida un tercer ojo en el rostro. El cerebro del animal se adaptó sin problemas a la nueva fuente de estímulos sensoriales.

Esto en cuanto a las herramientas. Respecto a las técnicas, sus efectos sobre el cerebro son idénticos. Los trabajadores que Maskelyne y De Prony contrataban para sus redes de calculadores humanos eran personas habituadas a operar con las técnicas de la matemática. La causa de aquel criterio de selección de personal es la misma, y es que la ejecución de los actos voluntarios mejora con la repetición (Ibíd., p. 491). Decía Aristóteles que el hábito es una especie de segunda naturaleza humana, y no se equivocaba, pues el hábito transforma la naturaleza biológica del cerebro en particular, y del cuerpo en general. La explicación neurológica de la modelación del cerebro por la repetición es el *principio de Hebb*, el cual establece que cuando una neurona A participa repetida o persistentemente en la excitación o inhibición de otra neurona B, entonces acontece algún tipo de proceso o cambio metabólico en una o ambas que incrementa la eficacia de A para excitar o inhibir a B (Ibíd., p. 681).

Por tanto, la tecnología modela nuestro cerebro. Los seres humanos creamos técnicas y herramientas como medios para obtener fines, pero esos medios no son neutrales, sino que nos afectan. La razón por la cual comúnmente se cree que son neutrales es porque ejercen su efecto sobre nosotros de manera silenciosa y gradual, de modo que no lo advertimos. Ésta es una consecuencia derivada de la fiabilidad de los utensilios o herramientas, que puede extenderse también a las técnicas. Como ya señalamos citando a Heidegger, la fiabilidad es la característica esencial de los utensilios. Son fiables para que podamos operar con ellos sin reparar en ellos. Mientras la parte consciente de nuestro pensamiento está ocupada en la acción a la que sirven como medio, ellos actúan no sólo *desde*, sino también *sobre* la parte inconsciente. Cuando nos hemos habituado a una tecnología, es tan difícil reparar en sus efectos sobre nuestra mente como difícil es el reparar en los efectos de nuestras manos sobre nuestra forma de comprender el mundo. Lo que siempre está ahí se

oculta a la mirada, tanto en el sentido figurado en el que lo afirma la fenomenología como en sentido literal. Cuando un estímulo actúa sobre nuestros órganos sensoriales sin variar su intensidad, cualidad o localización, al cabo de un cierto tiempo dejamos de percibirlo a causa de un proceso biológico a nivel de los canales iónicos neuronales denominado *desensibilización* (Ibíd., p. 260). Sólo percibimos lo que cambia. Lo que no cambia, no es. La neurociencia confirma así las intuiciones de Heráclito.

El carácter pedagógico, común a todas las tecnologías, es especialmente acusado en las de tipo intelectual, que son las que potencian nuestras facultades mentales, como es el caso de las computadoras electrónicas. Los efectos sobre la mente de las tecnologías intelectuales se conocen desde hace siglos. Platón, por boca de Sócrates, advierte en el *Fedro* de las consecuencias peligrosas de la escritura, la segunda de las grandes tecnologías intelectuales de la Historia, por detrás del lenguaje. Hacia el final del tercero de los discursos sobre el amor que componen el *Fedro*, Sócrates cuenta una leyenda sobre el origen de la escritura. La escritura, dice Sócrates, fue inventada por el dios Theuth. Como era costumbre en la época, Theuth acudió enseguida a presentarle su creación al entonces rey de Egipto, Ammón. La reacción del monarca fue la siguiente: «Oh, Theuth, excelso inventor de artes, unos son capaces de dar el ser a los inventos del arte, y otros de discernir en qué medida son ventajosos o perjudiciales para quienes van a hacer uso de ellos. Y ahora tú, como padre que eres de las letras, dijiste por cariño a ellas el efecto contrario al que producen. Pues este invento dará origen en las almas de quienes lo aprendan al olvido, por descuido del cultivo de la memoria, ya que los hombres, por culpa de su confianza en la escritura, serán traídos al recuerdo desde fuera, por unos caracteres ajenos a ellos, no desde dentro, por su propio esfuerzo. Así que no es un remedio para la memoria, sino para suscitar el recuerdo lo que es tu invento. Apariencia de sabiduría y no sabiduría verdadera procuras a tus discípulos. Pues habiendo oído hablar de muchas cosas sin instrucción, darán la impresión de conocer muchas cosas, a pesar de ser en su mayoría unos perfectos ignorantes; y serán fastidiosos de tratar, al haberse convertido, en vez de en sabios, en hombres con la presunción de serlo» (Platón, *Fedro*, 275a).

Hemos de aclarar que Platón se refiere a la escritura como un arte en vez de como una técnica debido a una decisión del traductor de la edición de la que hemos tomado la cita, pero en el original el término empleado es *tékne* (τέχνη), que con frecuencia se traduce por *ars*, "arte", y que es la raíz etimológica de "técnica" (Heidegger, 1952, p. 94). Por tanto, la técnica de la escritura tiene el efecto pedagógico de que aquellos que la usan corren el riesgo de desatender el cultivo de la memoria y de creer que saben lo que en realidad ignoran.

Las computadoras electrónicas pertenecen también a la categoría de las tecnologías intelectuales, y como tales su uso exige una serie de competencias cognitivas al tiempo que excluye otras. El uso de estas máquinas puede darse a dos niveles: ordinario y experto. El ordinario es el de la utilización de los programas informáticos o software, mientras que el experto es el de la creación de dichos programas. Vamos a ocuparnos sólo de este último tipo de uso, dado que las inteligencias artificiales, si es que son posibles sobre la base de las computadoras electrónicas, habrán de crearse elaborando programas informáticos. Las dos técnicas intelectuales más importantes requeridas en el proceso de programación de computadoras son los lenguajes formales y la escritura. Veámoslas.

Lenguaje formal

Las computadoras electrónicas son instanciaciones materiales de sistemas formales. La parte material es el hardware, y los sistemas formales que ejecuta son el software. Hay dos maneras de definir un sistema formal: mediante el *lenguaje natural* o mediante un *lenguaje artificial*. El ajedrez, por ejemplo, es un sistema formal que, como cualquier otro, puede ser definido mediante el lenguaje natural. También mediante el lenguaje natural se puede definir un conjunto de algoritmos cuya ejecución permitiría jugar bien al ajedrez. Con ese manual, un computador humano se convertiría en un ajedrecista de nivel tan excelente como Deep Blue. Sin embargo, si queremos que una computadora electrónica como Deep Blue ejecute un algoritmo, ya

sea para jugar al ajedrez o para cualquier otra finalidad, éste no puede serle proporcionado en un lenguaje natural, sino que debe serle suministrado en un lenguaje artificial de tipo formal. La diferencia entre un lenguaje artificial en general y un *lenguaje formal* es que este último está exento de interpretación semántica en su definición (Falguera & Martínez, 1999, p. 61).

Un lenguaje formal se compone de vocabulario y sintaxis. El *vocabulario* comprende un conjunto finito de expresiones o símbolos primitivos que, en el nivel más elemental de las computadoras, suele ser el binario de 1 y 0, mientras que la *sintaxis* consiste en un conjunto de reglas de formación que determinan qué secuencias de expresiones del vocabulario son fórmulas bien formadas. Así es como se escribían los programas de las primeras computadoras: para que el UNIVAC aprendiera a realizar una tarea algorítmica sencilla, como el cálculo de trayectorias balísticas, los ingenieros pasaban semanas escribiendo enunciados en símbolos binarios sobre un soporte material que habitualmente eran tarjetas perforadas o rollos de papel. En la actualidad, la programación es una tarea menos fatigosa gracias a la utilización de *lenguajes de programación de alto nivel* (*high-level programming languages*) (Deitel & Deitel, 2012, p. 10). La característica distintiva de éstos es que permiten introducir los algoritmos en un lenguaje formal de aspecto similar al natural, en vez de mediante secuencias de 1 y 0. Posteriormente, un programa especial denominado *compilador* (*compiler*) se encarga de traducir ese texto al lenguaje formal de nivel más elemental, también llamado *código máquina* (*machine code*), que es la serie de símbolos 1 y 0. Si el algoritmo ha sido bien escrito de acuerdo al vocabulario y la sintaxis del lenguaje de programación de alto nivel, entonces la máquina lo ejecuta.

Supongamos un sencillo algoritmo escrito en el lenguaje de alto nivel BASIC de Microsoft para averiguar si el número *x* es mayor o igual a 10. Se escribiría así: IF x >= 10 THEN PRINT "x es mayor o igual a 10" ELSE PRINT "x es menor de 10". Y se leería: Si x es mayor o igual a 10 entonces imprime en pantalla "x es mayor o igual a 10", de lo contrario, imprime en pantalla "x no es mayor o igual a 10". Se parece mucho al lenguaje natural que hablan los angloparlantes, pero no es igual. La diferencia es que

el lenguaje natural es flexible, mientras que los lenguajes de programación, en tanto que lenguajes formales, son extremadamente *rígidos*. Pongamos dos ejemplos de rigidez, uno relativo al vocabulario y otro relativo a la sintaxis. Rigidez en el vocabulario: si en vez de escribir THEN (entonces) el programador escribiera un sinónimo como THEREFORE (por tanto), la máquina no entendería esa instrucción, porque el compilador no sabe que, en el lenguaje natural y en ese contexto, THEN y THEREFORE quieren decir lo mismo. Su vocabulario es rígido en tanto que su extensión está perfectamente determinada, mientras que la extensión del vocabulario del lenguaje natural, aunque también es finita, es indeterminada. Rigidez en la sintaxis: si en vez de escribir x >= 10 el programador escribiera 10 <= x, la máquina tampoco lo entendería, porque las reglas de formación de enunciados de BASIC exigen que en una ecuación la variable (x) se escriba antes que la constante (10). Sin embargo, a un computador humano, que entiende el lenguaje natural, sí se le podría dar una instrucción que contuviese la expresión 10 <= x.

Una computadora electrónica, de acuerdo a la tesis Church-Turing, puede ejecutar cualquier algoritmo, pero sólo si se le proporciona en un lenguaje formal. Por el contrario, los seres humanos operamos la mayor parte del tiempo con lenguajes que pertenecen a la categoría del lenguaje natural. La diferencia básica entre ambos tipos de lenguaje es que los formales se construyen de manera artificial con reglas rígidas para garantizar la *univocidad* de los enunciados. Que son unívocos quiere decir que tienen un solo significado, mientras que los naturales están llenos de ambigüedades, paradojas y otros accidentes que dan lugar a la *equivocidad*, es decir, a la pluralidad de significados. La elección del significado de las expresiones equívocas depende de la pragmática, y la pragmática, como hemos mencionado antes, no ha sido formalizada algorítmicamente (Haugeland, 1981a, p. 28), y en el capítulo séptimo veremos por qué nunca lo será (Dreyfus, 1992, p. 198). Por tanto, las computadoras electrónicas, dado que sólo son capaces de ejecutar algoritmos, no podrían resolver el significado de enunciados equívocos. Aun suponiendo que la pragmática pudiera ser formalizada algorítmicamente para reducir mediante reglas los enunciados equívocos a otros

unívocos, en tal caso el código máquina resultante dependería del significado que el compilador hubiese interpretado que el programador "quería decir", lo cual supondría una importante pérdida de control sobre la máquina.

Por otra parte, existe en la sociedad una corriente positivista que considera a los lenguajes formales como formas de comunicación privilegiadas y que por tanto no sólo no deben ser sustituidos por lenguajes naturales, sino que, muy al contrario, deben ser ellos los que sustituyan a los lenguajes naturales. La supuesta superioridad de los lenguajes formales reclamada por el positivismo se debe a que son puramente operacionales. El *operacionalismo* es un modo de pensar propio de las ciencias físicas que reduce la descripción de los objetos a las operaciones mediante las cuales son obtenidos (Marcuse, 1964, p. 38). El significado unívoco referido por las proposiciones algorítmicas es justamente eso: una operación, porque los algoritmos no son más que operaciones para obtener fines. Para la computadora la variable x es unívoca porque sólo significa la operación para extraer de la memoria el valor asociado a ella. La mutilación del lenguaje natural para reducirlo a un lenguaje formal unívocamente operacional se observa con especial claridad en el ámbito de la programación informática, pero su alcance es universal y afecta a todas las esferas de la vida en nuestra época. Cualquier concepto extraño a dichas operaciones instrumentales es rechazado no sólo por el compilador de la computadora electrónica, sino también por la sociedad. Desde la Modernidad, la sociedad occidental lleva siglos modelándose a imagen y semejanza de las computadoras electrónicas, incluso antes de que éstas existieran (Weizenbaum, 1976, p. 9). En el mejor de los casos, las expresiones equívocas del lenguaje natural son toleradas dentro de una reserva especial del pensamiento. Así, es habitual oír hablar de verdad poética o verdad metafísica, como si el descubrimiento de la verdad a secas estuviera reservado a una metodología distinta de las de la poesía y la metafísica (Marcuse, 1964, p. 167).

La consecuencia práctica del carácter formal de los lenguajes de programación en lo que respecta a la creación del software es que se abre un abismo entre dichos lenguajes y el lenguaje natural que el programador utiliza para pensar durante la

mayor parte del tiempo. La distancia entre ambos sólo puede ser superada de dos maneras: adaptando a la máquina al modo de operar de la cognición humana, o adaptándose el ser humano al lenguaje formal con el que opera la máquina. La primera vía ha sido llevada a cabo con éxito al nivel del uso ordinario de los programas informáticos, pero no al nivel de su creación.

En la actualidad, el uso ordinario de las computadoras electrónicas es relativamente fácil para cualquiera gracias a la estandarización de los interfaces *WIMP*, acrónimo en inglés de ventanas (*windows*), iconos (*icons*), ratón (*mouse*) y menús desplegables (*pull-down menus*). WIMP fue inventado en el Centro de Investigación de Palo Alto de la compañía Xerox en los años 70, y posteriormente fue copiado y perfeccionado por Apple y Microsoft hasta convertirlo en lo que es hoy en día: la forma de comunicación o interfaz dominante entre el usuario corriente y la computadora electrónica. La facilidad para operar en entornos WIMP se debe a que éstos se adaptan a las habilidades cognitivas espontáneas del ser humano. Por ejemplo, para mover un archivo basta con pinchar sobre su icono con el ratón y arrastrarlo a la ventana destino. Lo aprendemos sin dificultad porque se trata de un procedimiento similar al que empleamos para mover un objeto físico: lo cogemos y, sin soltarlo, lo desplazamos hasta el lugar deseado.

Mientras que el uso ordinario de las computadoras se ha hecho más fácil en las últimas décadas gracias a los interfaces WIMP, la tarea de programarlas continúa siendo difícil debido a que para ello hay que utilizar lenguajes de programación. No parece plausible que un interfaz como WIMP, adaptado a las habilidades cognitivas espontáneas del ser humano, pueda implementarse a la programación. El motivo es que escribir algoritmos no guarda similitud con ningún procedimiento habitual para nuestra cognición, como mover objetos. De hecho, una de las mayores utilidades de la IA sería precisamente la de comprender las instrucciones de un ser humano enunciadas en un lenguaje natural y codificarlas en algoritmos. Por tanto, la única manera de programar ordenadores es, a día de hoy, adaptándose el ser humano a escribir en un lenguaje formal de programación.

La técnica para escribir un programa informático puede dividirse, a grandes rasgos, en dos fases. En la primera se diseña el algoritmo y se anota en un lenguaje mixto, o *pseudocódigo*, que combina el lenguaje natural con el lenguaje formal de programación (Deitel & Deitel, 2012, p. 103). En la segunda se traslada el algoritmo a la computadora formalizándolo del todo en el lenguaje de programación. Entonces llega el momento de ejecutar el compilador para que traduzca las instrucciones a código máquina, y pueden ocurrir dos cosas: que funcione o que no funcione. Si no funciona, es habitual recibir un mensaje que indica dónde se han cometido errores de sintaxis. En tal caso, corregirlos suele ser una tarea trivial. Por el contrario, si funciona, puede ocurrir que el ordenador haga lo que queríamos que hiciese, u otra cosa. Ésta última posibilidad significa que la segunda fase, la de formalizar el algoritmo en el lenguaje de programación, ha sido bien realizada, pero en la primera se ha cometido un error. Hallar ese error es una tarea asequible para el intelecto humano cuando se trata de algoritmos pequeños, pero cuando el algoritmo abarca miles de líneas, el programador puede pasar días, semanas o meses repasando línea por línea antes de dar con él, si es que tiene la fortuna de terminar encontrándolo. Este problema nos lleva a examinar la segunda técnica intelectual exigida para la programación de computadoras electrónicas: la escritura, con su pedagogía inherente.

Escritura

Los sistemas expertos, que son los deficientes geniales del reino de las máquinas, constan de muchas líneas de código, con una cantidad que puede oscilar entre las decenas de miles y los varios millones. Por tanto, es razonable suponer que las inteligencias artificiales en sentido fuerte, si algún día se consiguen mediante la programación de computadoras electrónicas, constarán también de varios millones de líneas de código. Es posible que surja un genio, un Einstein de la IA, que descubra un algoritmo que, con sólo diez o veinte mil líneas de código, sea capaz de una inteligencia tan plástica como la que produce el cerebro humano con sus cien mil

millones de neuronas y cien billones de conexiones. Sin embargo, lo razonable a día de hoy es suponer que las inteligencias artificiales en sentido fuerte serán programas por lo menos tan extensos como los sistemas expertos.

Por otra parte, un solo ser humano no podría diseñar y escribir un programa tan grande. Por tanto, las inteligencias artificiales en sentido fuerte tendrán que ser programadas por equipos de ingenieros que, previsiblemente, dividirán la inteligencia en módulos. Tal división modular puede ser por *funciones*, del tipo de aprender, planificar y razonar, o bien por *actividades*, desde las más simples como moverse hasta las más complejas como el lenguaje, para posteriormente integrarlas en una unidad (Brooks, 1991, p. 402). De hecho, así es como proceden los ingenieros para crear grandes programas informáticos: dividen las dificultades en otras más pequeñas. O, como dice Descartes: «Todo el método consiste en el orden y la disposición de los objetos sobre los cuales hay que centrar la penetración de la inteligencia para descubrir alguna verdad. Nos mantendremos cuidadosamente fieles a él si reducimos gradualmente las proposiciones complicadas y oscuras a proposiciones más simples, y luego, si partiendo de la intuición de las que son las más simples de todas, procuramos elevarnos por los mismos escalones o grados al conocimiento de todas las demás» (Descartes, *Reglas para la dirección de la mente*, p. 165).

El método cartesiano es la técnica ideal para crear programas informáticos. Su aplicación permite la comprensión de varias líneas de código de manera tan clara y distinta como si fueran una sola: «Las he de recorrer un cierto número de veces por medio de una especie de movimiento continuo de la imaginación que intuye de una sola mirada cada objeto en particular al mismo tiempo que pasa a los demás, hasta que haya aprendido a pasar de la primera proposición a la última con bastante rapidez como para que, sin dejar apenas ningún papel a la memoria, me parezca ver el todo de una vez por intuición» (Ibíd., p. 174). El inconveniente de esta técnica es que tiene un límite. Podemos comenzar, como propone Descartes, recorriendo con el pensamiento una proposición algorítmica simple, hasta comprenderla de manera intuitiva, perfecta. Y después dos, tres, cuatro.

Pero llega un momento en el cual es imposible añadir más por la sencilla razón de que el ser humano, además de ser una sustancia pensante, es una sustancia física que necesita descansar. Tras la interrupción, se hace necesario recurrir a la memoria de lo comprendido anteriormente, pues si se reiniciara el proceso jamás se podría pasar de las primeras líneas. Pero la memoria es falible, y por eso Descartes pretende evitar su participación: «Sin dejar apenas ningún papel a la memoria» (Ibíd., p. 174), dice. La mejor tecnología de la que dispone el ingeniero para remediar la falibilidad de su memoria es la escritura. Entre las líneas de código que deben ser ejecutadas por la computadora electrónica, se suelen introducir líneas con anotaciones que no serán ejecutadas, sino que describen en lenguaje ordinario las funciones de lo que está escrito en el lenguaje formal de programación. Cuanto más extenso y complejo es un programa, más recomendable es esta práctica, llegando a un punto en el que se hace indispensable. Los grandes proyectos, acometidos por decenas de ingenieros, implican necesariamente la redacción de informes o "memorias".

Los informes redactados por uno mismo son un buen remedio para la debilidad de la memoria, como proclamaba Theuth, pero cuando uno lee lo escrito por otros puede caer en el error de creer que sabe lo que en el fondo ignora, como replicaba Ammón (Platón, *Fedro*, 275a). Ya mencionamos antes que el sistema operativo Windows Vista, publicado en 2007, consta de 50 millones de líneas de código, obviamente escritas por muchos ingenieros. Durante su desarrollo, cada ingeniero trabajaba en un módulo del código, el cual posteriormente debía encajar con el resto. La forma de conseguir la armonía funcional en los grandes proyectos de software es mediante el intercambio de resúmenes que sintetizan lo que hace cada módulo. Es decir: se confía en que esos escritos permitan a cada una de las partes comprender lo que están haciendo las demás sin haber visto ni una sola de las líneas de código ajenas. El resultado de esta temeraria práctica es el que advertía Platón, y es que al final la totalidad del programa nunca funciona del todo bien. Las consecuencias son de sobra conocidas por cualquier usuario: el sistema se cuelga, se reinicia, muestra un mensaje de error, y, en definitiva, no hace lo que se suponía que debía hacer. Es algo similar a lo

que Descartes señala que sucede cuando una ciudad es diseñada por varios arquitectos: la obra final carece de un diseño único, coherente, armonioso (Descartes, *Discurso del método*, p. 53).

El fundador de la cibernética, Norbert Wiener, advirtió ya en los años 60 de que este fenómeno de la pérdida del control sobre lo que hacen las computadoras electrónicas podía suceder. Hace tiempo que estas máquinas han sobrepasado la comprensión de los seres humanos, y eso, según Joseph Weizenbaum, ha dado lugar a dos importantes consecuencias. La primera es que se toman decisiones basándose en programas informáticos que nadie comprende en su totalidad. Y la segunda es que dichos programas se vuelven inmunes al cambio, «porque en ausencia de una minuciosa comprensión del funcionamiento interior del sistema, cualquier modificación sustancial en él es muy probable que haga que todo el sistema se vuelva inoperativo y fácilmente irreparable. Estos sistemas de ordenador sólo pueden, pues, desarrollarse» (Weizenbaum, 1976, p. 195). El desarrollo consiste habitualmente en la adición de nuevas funciones o en parches para solucionar los problemas detectados. En ambos casos, el programa crece como la ciudad de Descartes, llena de callejuelas, vericuetos y caminos sin salida que lo convierten en un galimatías (Descartes, *Discurso del método*, p. 53). Windows 7, lanzado en 2009, fue construido sobre la base de Windows Vista. Los ingenieros de Microsoft, la compañía desarrolladora de ambos, se cuentan entre los mejores del mundo, pero ni siquiera ellos pueden superar el problema de la escritura descubierto por Platón en el siglo IV a.C., porque es, sencillamente, insalvable.

Resumen

Las computadoras electrónicas de propósito general, como por ejemplo los ordenadores personales, son formalmente equivalentes a la máquina universal de Turing, la cual, según la tesis Church-Turing, es capaz de efectuar cualquier algoritmo. Hay muchos resultados que pueden ser obtenidos mediante algoritmos, desde una

simulación virtual del juego de ajedrez hasta una IA capaz de jugarlo a gran nivel. Sin embargo, también hay resultados que no son obtenibles algorítmicamente, como demostró el propio Turing con el argumento de la secuencia diagonal. Junto con esta limitación formal, hemos señalado que las computadoras tienen limitaciones materiales. Por un lado, aunque su fiabilidad es tan alta que casi se pueden considerar digitales, no por ello dejan de ser falibles como cualquier otra herramienta, y por otro, el espectacular aumento de las capacidades para procesar información que han experimentado durante las últimas décadas parece estar ralentizándose por haber topado con los límites físicos de la miniaturización.

De cara a utilizar estas máquinas para la creación de inteligencias artificiales en sentido fuerte, aun suponiendo que sus limitaciones formales y materiales no fueran un obstáculo, todavía quedaría un gran problema que los investigadores no podrían resolver: ellos mismos, con sus limitaciones intelectuales para operar con lenguajes formales y para retener de manera exacta grandes volúmenes de información lógico-matemática codificada en dichos lenguajes. Los seres humanos, tal y como pensaba el matemático alemán David Hilbert, somos una fuente de error que, para ciertos fines, debe ser eliminada. Sin embargo, como demostraron Gödel, Turing y Church, somos imprescindibles para la tarea de acabar con nosotros mismos reemplazándonos por inteligencias artificiales. La cuarta punta de la estrella de Robinet indicaba esto: «Se necesita al Padre para completar la muerte del Padre» (Robinet, 1973, p. 48).

4. Contexto científico de la IA

La invención de las computadoras electrónicas programables al término de la Segunda Guerra Mundial fue un factor decisivo para que poco después, en la década de 1950, surgiera el cognitivismo, un paradigma interdisciplinar que se mantiene vigente en nuestros días y que se caracteriza por considerar que la mente es un procesador de información semejante a una computadora. En el seno del cognitivismo se desarrolló el primer programa de investigación de la IA (Franklin, 1995, p. 378), conocido como *IA simbólica*. Tras veinte años de andadura, y muchos millones de dólares invertidos por las agencias militares de los países desarrollados, a mediados de los 70 la IA simbólica se reveló incapaz de cumplir sus promesas, razón por la cual a comienzos de la década de los 80 fue desplazada por la *IA subsimbólica* o *conexionismo*, un programa de investigación alternativo que resurgió tras décadas de abandono y que se basa en el enfoque fisicalista de la neurociencia.

En la actualidad ambos programas coexisten, e incluso se combinan, siendo competidores en la carrera por construir la primera inteligencia artificial en sentido fuerte. La IA simbólica pretende hacerlo programando una computadora electrónica para que manipule símbolos de manera similar a como se supone que lo hace la *mente* según el paradigma cognitivista, mientras que la IA subsimbólica aspira a lograrlo reproduciendo o simulando el funcionamiento de las redes de neuronas del *cerebro*. La mente y el cerebro, dos enfoques opuestos pero complementarios para abordar la inteligencia. Este capítulo vamos a dedicarlo a examinar los conceptos fundamentales de las ciencias en las que se basan la IA simbólica y la IA subsimbólica: la psicología cognitiva y la neurociencia, respectivamente.

El psicólogo norteamericano Howard Gardner define *la ciencia cognitiva*, en singular, como «un empeño contemporáneo de base empírica por responder a interrogantes epistemológicos de antigua data, en particular los vinculados a la naturaleza del conocimiento, sus elementos componentes, sus fuentes, evolución y difusión» (Gardner, 1985, p. 21). Sin embargo, no es correcto hablar de la ciencia cognitiva en singular, dado que no existe una sola ciencia cognitiva, sino una pluralidad de ciencias que se denominan cognitivas en virtud de que comparten el cognitivismo como paradigma transversal. La noción de *paradigma* fue introducida en la filosofía de la ciencia por Thomas Kuhn en su obra de 1962 *La estructura de las revoluciones científicas*. Posteriormente, en una posdata incluida en la edición de 1970, la modificó para distinguir en ella un sentido general al que denominó *matriz disciplinar* y otro estricto al que dio el nombre de *ejemplar* (Kuhn, 1970, p. 175). Utilizaremos el término en su sentido general, el cual, como reconoce Kuhn, es por su propia naturaleza difícil de precisar. No obstante, podemos definirlo aproximativamente estableciendo que un paradigma es una estructura constituida por los supuestos teóricos generales, las leyes y las técnicas para su aplicación que son adoptados durante un tiempo por una determinada comunidad científica (Chalmers, 1982, p. 128).

Las ciencias cognitivas son aquellas que comparten el *paradigma cognitivista* (Carpintero, 1996, p. 404). Éste consta de dos supuestos teóricos fundamentales (García, 2001a, p. 18). El primero es la *metáfora computacional*, que consiste en la mencionada analogía según la cual la mente es un procesador de información semejante a una computadora electrónica. El segundo es la *tesis internalista* que establece la existencia de representaciones mentales y que reclama un nivel de análisis propio para estudiarlas al margen de los factores biológicos y de algunos de los factores ambientales que las afectan. En cuanto a las leyes del cognitivismo, pueden dividirse en dos grupos. Por un lado, las que han sido descubiertas durante los últimos cincuenta años dentro del paradigma cognitivista, y por otro, todas aquellas leyes de la psicología que hayan sido formuladas dentro de otros paradigmas siempre y cuando se adecúen a sus dos supuestos teóricos fundamentales. Y, finalmente, las técnicas del

cognitivismo proceden también en algunos casos de etapas anteriores, como por ejemplo la psicometría, fundada a finales del siglo XIX por Francis Galton y Alfred Binet, pero las más características son las nuevas tecnologías de base informática, tales como la electroencefalografía (EEG), la tomografía computarizada por emisión de positrones (PET), la resonancia magnética funcional (fMRI), la magnetoencefalografía (MEG) y la tomografía computarizada por emisión de fotones simples (SPECT).

Otra manera clásica de estructurar las teorías científicas, aparte de la de Thomas Kuhn, es la propuesta por Imre Lakatos. El equivalente al concepto kuhniano de paradigma en el texto de Lakatos es el concepto de *programa de investigación*. Según Lakatos, las teorías científicas se articulan en programas de investigación que sirven para guiar la investigación futura. Dentro de los programas se distinguen dos zonas: el núcleo central y el cinturón protector. El *núcleo central* reúne las proposiciones que constituyen la esencia del programa, mientras que el *cinturón protector* se compone de todo lo demás, ya sean leyes, hipótesis o técnicas. La actividad de una comunidad científica adherida a un determinado programa de investigación consiste en realizar dos tipos de *heurística*: una negativa y otra positiva. La heurística *negativa* es el compromiso de defender las proposiciones del núcleo central, al tiempo que la *positiva* se concreta en la ampliación progresiva del núcleo central mediante la consolidación de las proposiciones del cinturón. Siguiendo con la equivalencia entre las teorías de Kuhn y Lakatos, el núcleo central de un programa de investigación correspondería a los supuestos teóricos generales de un paradigma, mientras que el cinturón protector abarcaría las leyes y las técnicas.

Ciertamente, hay notables diferencias entre las teorías de Kuhn y Lakatos, pero residen sobre todo en la importancia desigual que conceden a los factores sociológicos dentro de la ciencia, siendo considerable en el caso de Kuhn, y escasa para Lakatos (Chalmers, 1982, p. 128). Si dejamos al margen los factores sociológicos que afectan a la justificación de las teorías científicas, entonces no hay objeciones que impidan establecer, tal y como haremos de aquí en adelante, una correspondencia entre los paradigmas de Kuhn y los programas de investigación de Lakatos.

4.1. Conceptos de filosofía de la ciencia

Para cumplir con el propósito de analizar en este capítulo las ciencias en las que se basan los programas de investigación simbólico y subsimbólico de la IA, es necesario que dediquemos una sección preliminar a exponer varios conceptos fundamentales de filosofía de la ciencia. Algunos los utilizaremos de inmediato, y otros nos servirán más adelante. Vamos a comenzar por definir qué es la ciencia misma. En cuanto a las pretensiones epistemológicas de las teorías científicas vamos a describir los enfoques *realista* e *instrumental*; respecto a la naturaleza de los factores que afectan a los procesos de creación y justificación de dichas teorías expondremos los puntos de vista del *racionalismo* y del *relativismo*; y finalmente expondremos las diferencias entre la *explicación* de los hechos naturales y la *comprensión* de los hechos sociales. Aclaramos de antemano, tal y como hicimos en el capítulo segundo a propósito del racionalismo y del empirismo, que dentro de cada una de las tendencias que vamos a exponer hay diferencias entre los autores que las defienden. No obstante, nuestro cometido no es ahondar en tales discrepancias, sino ofrecer una panorámica general que sirva a nuestros intereses focalizados en la IA.

Realismo e instrumentalismo

El problema de definir la ciencia se conoce como el *problema de la demarcación*. Los intentos por resolverlo han consistido tradicionalmente en tentativas de definir la ciencia como el conocimiento que resulta de aplicar el método científico. Si el método científico existiera, proporcionaría una definición precisa de al menos uno de los dos contextos principales de la ciencia. Estos son el *contexto de descubrimiento*, que hace referencia a los factores que participan en la creación de las teorías científicas, y el *contexto de justificación*, que explica el modo en que dichas teorías son validadas. Este último puede ser concebido en términos exclusivamente lógicos, como

hacen los partidarios del *racionalismo* de la ciencia al estilo de Lakatos, o bien admitiendo el concurso de elementos sociológicos y en general psicológicos, como hacen los defensores del *relativismo* en la línea de Kuhn. Si existiera un método exclusivo de la ciencia para descubrir o justificar el conocimiento, entonces el problema de la demarcación quedaría resuelto. La cuestión es que no existe un tal método científico. Se habla continuamente de él como si existiera, pero en realidad no es más que una ilusión producida por la repetición de una mentira.

No obstante, aunque la ciencia no haya sido bien definida todavía, lo cierto es que posee rasgos esenciales que nadie pone en duda y que sirven de guía a los intentos por capturar su definición. Uno de esos rasgos es que las teorías científicas han de ser *válidas*. Parece una obviedad, pero no lo es si tenemos en cuenta que la validez puede ser entendida al menos desde dos puntos de vista: realismo e instrumentalismo. Los partidarios del *realismo* sostienen que la validez de una teoría depende del grado en que ésta sea verdadera. Cuanto más verdadera sea una teoría, más se parecerá a la Idea platónica de la que es imitación, y por tanto en mayor grado poseerá los atributos de inmutabilidad, eternidad y universalidad. En cambio, los defensores del *instrumentalismo* no están obligados a asumir ningún compromiso ontológico ni epistemológico acerca de la verdad. Su argumento es que el objetivo de la ciencia no es elaborar teorías verdaderas, sino exitosas, y el éxito depende no sólo de la fiabilidad proporcionada por los atributos de inmutabilidad, eternidad y universalidad, sino también de la utilidad, y la utilidad es un concepto social. Lo que en unas circunstancias o contexto cultural es útil, en otras puede no serlo. Por ejemplo, la mecánica de Newton se ha demostrado que no es verdadera, pero sigue siendo válida gracias a que es útil para operar a escala mesocósmica en condiciones no relativistas. De hecho, es la mecánica que utilizan los ingenieros aeroespaciales para poner satélites en órbita. Podrían hacerlo utilizando la teoría de la relatividad de Einstein, pero para este propósito la teoría de Newton es más cómoda. De manera análoga, una cuchara es más precisa que un cubo, pero para vaciar una charca el cubo es mejor, más eficiente que la cuchara.

Karl Popper, uno de los más ilustres representantes del realismo científico, está convencido de que la falsación es un proceso que conduce a la obtención de teorías cada vez más cercanas a la verdad sobre cómo funciona el universo (Chalmers, 1982, p. 218). Según Popper, cada vez que se demuestra que una teoría es falsa, la siguiente que la sustituye está más cerca de la verdad, y así sucesivamente en un proceso quizás asintótico sin final. Contra esta opinión los instrumentalistas esgrimen un argumento que, a nuestro entender, es concluyente, y por tanto suscribimos. Tomemos la mecánica de Newton y la teoría de la relatividad como ejemplos de teorías rivales que compiten por explicar la misma parcela de fenómenos del universo. Los realistas como Popper afirman que la teoría de la relatividad es más verdadera que la mecánica newtoniana, entendiendo la verdad en el sentido clásico de copia o correspondencia: *adequatio rei et intellectus*. Así, una proposición es más verdadera cuanto más se parezca el estado de cosas por ella descrito a la realidad en sí misma.

El problema es que para determinar el grado de semejanza de dos teorías respecto de la realidad en sí misma a la que se refieren es condición indispensable conocer la realidad en sí misma, pero éste no es el caso y quizás hasta sea imposible. Por tanto, dado que se desconoce el referente de la comparación, no hay manera de decidir cuál es la teoría más verdadera, si la de Newton o la de Einstein. Retomando la metáfora, sólo sabemos que la una es un cubo y la otra una cuchara, pero no sabemos si la verdadera mecánica se parece más a un cubo o a una cuchara. Además, podría suceder incluso que las verdaderas leyes del universo no sean expresables en forma de proposiciones matemáticas. A este respecto, comentando el análisis de Husserl de la estructura socio-histórica de la razón científica, Marcuse señala que las nociones de la exactitud y la fungibilidad matemática en las que se basa la ciencia moderna no garantizan la verdad objetiva, sino que «envuelven una experiencia concreta específica de la *Lebenswelt*: un modo específico de "ver" el mundo» (Marcuse, 1964, p. 151). La ciencia moderna pertenece, por tanto, a un proyecto histórico respecto al cual sus productos tienen una validez relativa. Así, si ese proyecto histórico cambiase en favor de otro, tal y como en su día se pasó de la física aristotélica a la galileana, entonces no

tendría sentido plantearse si Einstein estuvo más cerca de la verdad que Newton, porque ambos se habrían basado en un supuesto metateórico falso: el de la naturaleza matemática de las leyes del universo. Como tampoco tendría sentido plantearse tal comparación si se descubriese, en contra del dogma positivista, que las leyes del universo cambian con el paso del tiempo.

Los realistas posteriores a Popper son conscientes de esta debilidad de su postura, por lo que algunos la han modificado hacia un tipo de *realismo no representativo*. Según Alan Chalmers, «el realismo no representativo *no es representativo* en la medida en que no conlleva una teoría de la verdad como correspondencia. El realista no representativo no supone que las teorías describan entidades del mundo, tales como ondas, funciones o campos, en la forma en que nuestras ideas propias del sentido común entienden o nuestro lenguaje describe las mesas y los gatos» (Chalmers, 1982, p. 226). Una vez que se ha renunciado a la verdad como copia, es necesario reemplazarla por algún otro criterio que sirva para dirimir el grado de validez de las teorías científicas y, por tanto, para dar cuenta de las causas que deciden la sustitución de unos paradigmas por otros cuando acontece una revolución científica. Ese criterio, dice Chalmers, es el éxito: «Podemos juzgar nuestras teorías desde un punto de vista como el grado en que abordan con éxito algún aspecto del mundo» (Ibíd., p. 226). Sin embargo, el éxito es un concepto subjetivo, determinado por factores ajenos a la propia teoría que dependen del uso que quiera darle el agente que la utiliza. Por tanto, en el momento en que un realista como Chalmers se da cuenta de que debe renunciar a la verdad como adecuación, se ve obligado a pasarse a un enfoque instrumental de la ciencia.

Así pues, las teorías científicas han de ser concebidas, a nuestro juicio, como constructos intelectuales cuya validez depende de su éxito para lograr fines. Si recuperamos la definición de técnica que formulamos al final del capítulo anterior, según la cual la técnica es «toda serie de reglas por medio de las cuales se consigue algo» (Ferrater, 1965b, p. 763), llegamos entonces a la conclusión ineludible de que la ciencia no es más que una técnica evolutivamente superior (Habermas, 1968a, p. 194),

producto por tanto de la razón instrumental. Los contextos de descubrimiento y justificación de una tecnología, como por ejemplo el martillo, son exactamente iguales a los de una teoría científica. En ambos casos participan factores lógicos y psicológicos. La ciencia, en este sentido, no tiene nada de especial. Es una esfera de la actividad humana como cualquier otra, perteneciente a un proyecto histórico con un modo particular de ver el mundo. Lo único que podría hacerla especial sería el estar basada en un método. Pero, como ya hemos adelantado y vamos a argumentar a continuación, el método científico no existe, sino que es un mito generado por la repetición de una mentira, igual que la IA fuerte.

Racionalismo y relativismo

Un *método*, atendiendo a la etimología del término, es un "camino para alcanzar una meta". En este sentido, y sólo en éste, afirmamos rotundamente que el método científico es un mito y no existe, dado que no hay ningún camino, ni en el contexto de descubrimiento ni en el de justificación, que lleve a la obtención de teorías científicas: «En un sentido filosóficamente neutral de *ciencia*, el término *método científico* supone la existencia de unos procedimientos cuya aplicación garantiza el logro de conocimiento» (Rivadulla, 2009, p. 231). Lo que sí existe es una serie de características genéricas de la ciencia, de las que hablaremos en el apartado siguiente, y que son diferentes para las ciencias naturales y las ciencias sociales. Pero dichas características no son constitutivas de sendos métodos, sino más bien de modelos de comportamiento, de modos de proceder con respecto al estudio de la naturaleza y de la sociedad. Esos modos de proceder, que indican cómo avanzar pero no generan el avance, reciben el impulso hacia la meta del conocimiento no de un método, sino de diversas *estrategias* de naturaleza *lógica* y *psicológica* (Ibíd., p. 231), que no garantizan llegar al destino, en tanto que son falibles, y que además son en su mayoría comunes a todo tipo de conocimiento, ya sea natural o social, científico o precientífico. Veamos ambos tipos de estrategias en los dos contextos principales de la ciencia.

En cuanto al contexto de descubrimiento, la creatividad científica emplea al menos tres estrategias lógicas y dos psicológicas (Ibíd., p. 234). Las estrategias lógicas son la inducción, la abducción y la preducción. La *inducción* consiste en generalizar una ley universal a partir de una lista finita de enunciados observacionales. Sus exigencias son que el número de enunciados observacionales sea grande, que las observaciones se repitan en una amplia variedad de condiciones y que ningún enunciado observacional entre en contradicción con la ley universal (Chalmers, 1982, p. 15). Como bien se sabe desde Aristóteles, la inducción no es un método válido de producción de conocimiento debido a las deficiencias lógicas contenidas en todas sus exigencias. En primer lugar, porque por muy grande que sea el número de observaciones realizadas, siempre es ínfimo en relación a la cantidad de observaciones potenciales. En segundo, porque el número de condiciones que participan en la observación de un fenómeno también son infinitas, y por tanto no pueden examinarse todas. Y, finalmente, porque el hecho de que un número finito de enunciados observacionales no contradiga la ley universal no garantiza que en el futuro no vaya a aparecer un nuevo enunciado que sí sea contradictorio.

Hay una razón todavía más sutil en contra de la inducción, y es que todo enunciado observacional implica una carga teórica de cuya validez depende la suya. Por ejemplo, la observación a través de un telescopio supone que las leyes de la óptica sobre las que ha sido construido el aparato son válidas. O, como dice Steve Woolgar, para medir la temperatura se emplean termómetros cuyas lecturas dependen de la validez de las teorías sobre la dilatación del mercurio (Woolgar, 1988, p. 134). E incluso sin instrumentos, la observación de que un cisne es blanco presupone que a la luz del día el color blanco es percibido sin distorsiones cromáticas significativas por nuestro aparato visual. Los enunciados observacionales son tan falibles como las teorías que presuponen. Por tanto, la inducción no es un método válido para generar proposiciones científicas ni para justificarlas, pero sí es una estrategia que los científicos utilizan de hecho para proponer hipótesis, y que asimismo utilizamos todos los seres humanos de manera precientífica en nuestra vida ordinaria.

La segunda estrategia lógica utilizada en el contexto de descubrimiento es la *abducción*. Tal como la definió Charles S. Peirce, la abducción consiste en estudiar hechos e inventar una teoría que los explique (Rivadulla, 2010, p. 121). En este proceso el científico observa un hecho C. A continuación se le ocurre que si la teoría A fuera verdadera entonces C quedaría explicado. Y concluye que hay razón para sospechar que A es verdadera. El primer problema que encontramos en la abducción es cómo surge la teoría A en el segundo paso. A este respecto, Albert Einstein, que algo sabía de crear teorías científicas, afirma taxativamente: «No hay camino lógico que lleve a estas leyes fundamentales. Debemos dejarnos conducir por la intuición, que se basa en una sensación de la experiencia. [...] Nadie que haya profundizado de veras en esto podrá negar que el sistema teórico ha sido prácticamente determinado por el mundo de las suposiciones, pese a que no existe camino lógico alguno que conduzca desde éstas hasta las leyes fundamentales» (Einstein, 1955, p. 131). El propósito del método de Descartes era definir ese camino lógico que, como dice Einstein, en realidad no existe. En este sentido, la abducción puede considerarse como una estrategia psicológica, pero la clasificamos como lógica en virtud del silogismo formado por sus tres pasos. Dicho silogismo es una *falacia de la afirmación del consecuente*, según la cual: $[C \wedge (A \rightarrow C)] \rightarrow A$. Realizando la correspondiente tabla de verdad se descubre que no es una tautología, es decir, que es un razonamiento que no siempre es verdadero. Por tanto, la abducción tampoco es un método para generar conocimiento. Al igual que la inducción, es sólo una estrategia falible para producir hipótesis.

Finalmente, la *preducción* consiste en aplicar la deducción, que típicamente es una estrategia del contexto de justificación, al contexto de descubrimiento. A diferencia de la abducción, el punto de partida no son enunciados observacionales, sino teorías ya conocidas. La preducción es la forma de razonamiento que consiste en recurrir a los principios aceptados, a fin de anticipar propuestas teóricas por medio de la combinación y manipulación matemáticas, compatibles con el análisis dimensional, de los principios empleados (Rivadulla, 2010, p. 120). En consecuencia, desde un punto de vista lógico la preducción no amplía el conocimiento disponible, dado que las

conclusiones ya estaban contenidas en las premisas, pero sí lo hace desde un punto de vista epistemológico, en tanto que las ecuaciones generadas revelan nuevos enfoques. Una proposición generada preductivamente es tan falible como lo sean las teorías de las cuales ha sido derivada. Gracias a esto, la preducción es una estrategia bastante sólida, pero tiene el inconveniente de que sólo puede aplicarse a las ciencias teóricas, como la física, cuyo estilo explicativo es el *nomológico-deductivo*, es decir, el estilo explicativo que consiste en expresar las regularidades de los fenómenos en forma de relaciones ecuacionales entre parámetros cuantitativos para posteriormente deducir consecuencias particulares sustituyendo las variables por valores concretos. A juicio de los positivistas, éste es, por su exactitud predictiva, el estilo explicativo modélico para todas las ciencias (Horkheimer, 1947, p. 102).

Respecto a las estrategias psicológicas del contexto de descubrimiento, son por lo menos dos: la serendipia y el razonamiento por analogía. La *serendipia* es un descubrimiento casual. El ejemplo más famoso quizás sea el de la penicilina, que fue descubierta accidentalmente por Alexander Fleming al observar que en una placa de cultivo descuidada que se disponía a destruir había crecido un hongo rodeado de unas misteriosas colonias bacterianas. Por su parte, el *razonamiento por analogía* es una forma de inducción: dado que el individuo A tiene el conjunto de propiedades P, y que B tiene buena parte de las propiedades P, se concluye que B debe de tenerlas todas. Ambas, serendipia y razonamiento por analogía, son estrategias falibles, pero utilizadas de hecho por los científicos, de manera consciente e inconsciente.

En conclusión, para producir nuevas teorías no existe un método científico que garantice su obtención, sino sólo una variedad de estrategias falibles de naturaleza lógica y psicológica. Nótese que nuestra crítica no es que las teorías resultantes sean falibles, pues deben serlo siempre, sino que el proceso para producirlas es falible. Y lo que es más importante: dichas estrategias son comunes tanto al conocimiento científico como al precientífico, y se utilizan tanto en las ciencias de la naturaleza como en las ciencias sociales, con la única excepción de la preducción, que es exclusiva de las ciencias naturales de tipo nomológico-deductivo. Ahora bien, si el método científico

consistiera en la preducción, entonces sería una herramienta de segundo orden, dado que opera sobre teorías que, en algún momento anterior, han tenido que ser descubiertas mediante otra clase de estrategia. Y, por otra parte, el elenco de ciencias quedaría reducido a aquellas cuyo estilo explicativo es el nomológico-deductivo. De esta manera sólo las matemáticas, la física y la química serían disciplinas científicas. La geología, por ejemplo, quedaría excluida, dado que, además de utilizar explicaciones nomológico-deductivas gracias a su colaboración interdisciplinar con las matemáticas, la física y la química, también emplea explicaciones morfológicas y sistemáticas.

Las *explicaciones morfológicas* son aquellas que explican un fenómeno apelando a la estructura y a las propiedades de las partes que lo componen (Haugeland, 1978, p. 246). Una explicación morfológica tomada de la geología es la que da cuenta de que los sedimentos se acumulan en estratos en vez de mezclarse debido a que los cuerpos sólidos no se disuelven como los líquidos. Por su parte, las *explicaciones sistemáticas* son similares a las morfológicas, con la diferencia de que el fenómeno explicado surge de la interacción de las partes (Ibíd., p. 247). Un buen ejemplo es la explicación de cómo funciona un coche. La descripción morfológica de los componentes de un vehículo no es suficiente para explicar cómo funciona, sino que hace falta exponer el modo en que dichos componentes interactúan entre sí. En geología una explicación sistemática sería la teoría de la deriva continental, que explica el movimiento de los continentes como resultado de la interacción de las placas tectónicas con el magma y otros elementos.

Por tanto, a menos que se tenga la pretensión positivista extrema de reducir la ciencia a las matemáticas, la física y la química, no existe un método científico en el contexto de descubrimiento. E incluso en el caso de la preducción, exclusiva de estas tres ciencias, no es suficiente por sí sola para producir conocimiento, dado que necesita de teorías previamente obtenidas por otros medios. Todas las ciencias, ya sean naturales o sociales, emplean las mismas estrategias para generar conocimiento: inducción, abducción, serendipia y razonamiento por analogía. Ninguna garantiza que su aplicación producirá teorías científicas.

En cuanto al contexto de justificación, tampoco en él encontramos el método científico. Las teorías se justifican por procedimientos lógicos y psicológicos. Los procedimientos lógicos por excelencia son la falsación y la repetibilidad. Sobre la *falsación* Woolgar comenta: «Según Popper, la esencia de la metodología científica consiste en producir generalizaciones que resistan los intentos de falsación. Debería intentarse verificar las proposiciones que contradijesen a la generalización de que se tratase; el fracaso de la verificación de la contra-proposición (esto es, el fracaso de la falsación) daría credibilidad (cuando menos temporalmente) a dicha generalización» (Woolgar, 1988, p. 25). Contra las pretensiones de Popper, la verificación o, más correctamente, la *corroboración* mediante la resistencia a la falsación, no vale como criterio metodológico de demarcación de la ciencia por al menos dos razones. Primero, porque los enunciados observacionales destinados a falsar la generalización conllevan una carga teórica problemática, como ya hemos comentado al hablar de la inducción. Y segundo, porque hay numerosas teorías que han sido falsadas y que, sin embargo, son científicas. La mecánica de Newton, por ejemplo, ha sido falsada porque se ha demostrado que no explica ciertos fenómenos que debería explicar, pero nadie duda de que es una teoría científica. La razón por la que continúa siendo científica tras haber sido falsada es porque sigue siendo útil, un criterio instrumental y sociológico.

En cuanto a la *repetibilidad* de los fenómenos sobre los que se pronuncian las teorías, es una exigencia aplicable a las ciencias de la naturaleza *experimentales*, tales como la física y la química, pero no a las *observacionales* como la paleontología (Rivadulla, 2010, p. 122). Los enunciados de las ciencias experimentales han de ser repetibles, que es justo la característica que les falta a las pseudociencias como la parapsicología, mientras que las observacionales no pueden repetir las experiencias a las que se refieren porque no depende de la voluntad del investigador que se produzcan nuevos fenómenos que avalen su teoría, como por ejemplo, nuevos restos fósiles de un cierto tipo en un lugar concreto. Si la repetibilidad fuera la condición de cientificidad, entonces sólo las ciencias experimentales podrían ser científicas, algo que es contrario a la noción real de la ciencia, la que de hecho se utiliza.

Y, finalmente, respecto a los factores psicológicos que actúan en el contexto de justificación, ya hemos mencionado que son rechazados por aquellos que, como Popper y Lakatos, defienden una concepción racionalista de la ciencia, al tiempo que son aceptados por los relativistas como Kuhn. Nosotros nos posicionamos del lado de estos últimos, dado que un enfoque instrumental de la ciencia, como el que antes hemos defendido, va siempre ligado al relativismo. La razón de este nexo es que, al afirmar que las teorías científicas son instrumentos cuya validez depende de su grado de eficacia, y al ser la eficacia un concepto sociológico y por tanto, en última instancia, psicológico, la justificación de las teorías científicas depende de factores psicológicos. En consecuencia, los argumentos antes empleados en favor del instrumentalismo valen también aquí para justificar el relativismo.

Explicar y comprender

En el apartado anterior hemos justificado la tesis de que no existe un método científico, sino tan sólo una multiplicidad de estrategias falibles para producir conocimiento, las cuales, en su mayoría, son comunes al conocimiento científico y precientífico, y son compartidas por las ciencias naturales y sociales. Ahora bien, esto no quiere decir que ambas clases de ciencia deban confundirse en una sola. Muy al contrario, como ya hemos anticipado, existen diferencias, todas ellas referidas a los distintos modos de proceder en el estudio de sus respectivos objetos: naturaleza y sociedad. Jürgen Habermas señala varias (Habermas, 1968b, p. 169).

La primera es que el trascendental de las ciencias naturales, o *empírico-analíticas* como él las denomina, es la medición, es decir, el número, frente a la gramática, que es el de las ciencias sociales, o *histórico-hermenéuticas*. Las demás características de las ciencias naturales enumeradas por Habermas son que su interés es técnico-instrumental, su objeto son los hechos naturales, su sujeto es la comunidad de investigadores, su método es explicativo, su producto es la información, entienden la verdad como copia, su lenguaje es formal o por lo menos formalizable y su medio es

el trabajo. Respecto a las ciencias sociales, su interés es práctico-comunicativo, su objeto son los hechos humanos, su sujeto es la especie humana, su método es comprensivo, su producto es la interpretación, entienden la verdad como consenso y su lenguaje y su medio son el lenguaje ordinario.

Esta disparidad de características es un hecho que afecta singularmente a la psicología. La razón es que el objeto de estudio de la psicología es la mente, y la mente, como cualquier otro rasgo fenotípico, es el resultado de la interacción entre factores biológicos y ambientales. Así, la psicología es una disciplina *biosocial*. Por su condición de biológica recurre a los conocimientos proporcionados por algunas ciencias de la naturaleza, y por su condición de social necesita de las ciencias sociales. Esta mezcla de disciplinas heterogéneas, que por un lado utilizan un método explicativo y por otro un método comprensivo, es la causa de que la psicología carezca de unidad en torno a un solo paradigma. Existe una gran variedad paradigmas para el estudio de la mente sostenidos por diversas escuelas de psicología. Los supuestos teóricos, las leyes y las técnicas de cada uno de esos paradigmas dependen de la importancia relativa que cada escuela concede a las partes natural y social del ser humano. Explicar ambas en el núcleo central de un solo paradigma es, como mínimo, extremadamente difícil, y tal vez sea imposible. El *método explicativo* de las ciencias naturales consiste en un enfoque *molecular*, que va de lo particular a lo general, *de abajo a arriba* (*bottom-up*), para dar cuenta de los fenómenos complejos mediante su descomposición en otros más simples o atómicos. Por su parte, el *método comprensivo* de las ciencias sociales consiste en un enfoque *molar*, que va de lo general a lo particular, *de arriba a abajo* (*top-down*), para dar cuenta de los fenómenos complejos no mediante su descomposición, sino describiéndolos tal como se presentan y en su relación de circularidad hermenéutica con los fenómenos más simples.

Una ciencia verdaderamente biosocial, como aspira a serlo la psicología, tiene la obligación de integrar en un solo paradigma todas estas diferencias, pero el mero hecho de intentar una síntesis tan compleja supone un problema epistemológico de tal envergadura que las diversas escuelas de psicología tienden a evitarlo posicionándose

cerca de uno de los dos extremos. Las que se acercan más a la metodología de las ciencias naturales se dice que forman parte de la corriente *explicativa*, mientras que las más próximas a la metodología de las ciencias sociales pertenecen a la corriente *comprensiva*. Esta separación se conoce en psicología como el *problema del método*, el cual sigue pendiente de solución (García, 2001a, p. 118). Así, por ejemplo, las escuelas de orientación positivista, como los conductistas, intentan hacer psicología basándose en la observación de hechos cuantificables, como es propio de las ciencias de la naturaleza, frente a las escuelas humanistas, como es el caso de los psicoanalistas, que pretenden explicar la mente centrándose en los factores sociales. En cuanto a la psicología cognitiva, que es la que nos ocupa aquí por ser el paradigma en el que se inscribe el programa de investigación de la IA simbólica, tampoco se salva del problema del método, como vamos a ver a continuación. Comenzaremos por examinar sus orígenes retomando la narración de la Historia de la psicología donde la habíamos dejado, en la disputa del siglo XVII entre empiristas y racionalistas.

4.2. Psicología cognitiva

Como vimos en el capítulo segundo, las opiniones que los empiristas y los racionalistas tenían sobre la mente eran muy distintas. Sin embargo, todas ellas procedían de una misma fuente de información: el método introspectivo. La *introspección* consiste en la inspección que hace el sujeto de sus propios actos psíquicos. Sus deficiencias fundamentales son dos. La primera es que no alcanza a dar cuenta de los fenómenos mentales inconscientes, dado que por definición acontecen sin que el sujeto se aperciba de ellos. Y la segunda es que, como decía Hans-Georg Gadamer, la lente de la subjetividad es un espejo deformante (Gadamer, 1960), por lo que el relato que el sujeto hace de un fenómeno psíquico interno no es un enunciado observacional, sino otro fenómeno psíquico basado en el anterior. Prueba de ello es que sirvió al mismo tiempo para argumentar en favor de teorías psicológicas tan opuestas como las de Descartes y Hobbes.

A pesar de su manifiesta incapacidad para constituir al menos por sí sola una fuente fiable de conocimiento, los investigadores de la mente continuaron confiando en la introspección hasta inicios del siglo XX. Tal fue el caso de Wilhelm Wundt. En la segunda mitad del siglo XIX, Wundt refundó la psicología como ciencia positiva. Su crucial aportación consistió en separarla de la filosofía reclamando para ella la condición de disciplina científica con métodos, programas, e instituciones propias (Gardner, 1985, p. 120). Sin embargo, la edición de revistas especializadas y la organización de congresos no suelen ser condiciones suficientes para constituir una ciencia. Wundt cometió el error de seguir utilizando la introspección como fuente de información, razón por la cual no logró establecer en la psicología un paradigma sólido y estable en torno al cual pudieran agruparse todos los investigadores para hacerla avanzar desarrollando su heurística positiva.

No se hicieron esperar las críticas procedentes de numerosas corrientes de la psicología de la época, tales como la escuela de Wurtzburgo dirigida por Oswald Külpe, los funcionalistas anglosajones con William James a la cabeza y la escuela reflexológica rusa con Ivan Pavlov como máximo exponente. Estas dos últimas corrientes, el funcionalismo y la reflexología, son las que más nos interesan aquí, dado que pusieron algunas de las bases principales para la aparición del conductismo poco después, a principios del siglo XX. Por su parte, los funcionalistas como William James, James Angell y John Dewey se opusieron al estructuralismo de Wundt porque sostenían que la vida psíquica debía ser entendida primordialmente no como una estructura, sino desde un punto de vista evolucionista y pragmático según el cual se trata de una herramienta que se ha consolidado en algunas especies porque resulta útil para la supervivencia y la satisfacción de los deseos del individuo.

En cuanto a la escuela reflexológica rusa, sus pioneros fueron Ivan Sechenov e Ivan Pavlov. Como médico que era, Sechenov se propuso demostrar que todos los procesos mentales tienen una base fisiológica y consisten en una actividad refleja, ya sea innata o aprendida. Es fácil apreciar que el monismo materialista de esta pretensión guarda continuidad con el pensamiento de los empiristas corporalistas

como Hobbes. El esquema del *arco reflejo*, compuesto por el par S-R de estímulo y respuesta, era en opinión de Sechenov suficiente para explicar toda la conducta, tanto involuntaria como voluntaria. Por su parte, Pavlov realizó uno de los hallazgos más importantes en la Historia de la psicología: la *ley del condicionamiento clásico*. Existen dos tipos de estímulos, los condicionados y los incondicionados, en función de la manera en la que elicitan una respuesta. Los incondicionados llevan una respuesta asociada de manera natural, no aprendida, como es el caso de la comida, cuya presencia produce salivación, mientras que los condicionados dan lugar a una respuesta sólo después de un aprendizaje. Pavlov descubrió que un estímulo condicionado que se presenta repetidamente junto a otro incondicionado acaba por desencadenar la respuesta asociada de manera natural al estímulo incondicionado. Ésta es la ley del condicionamiento clásico. El famoso perro de Pavlov fue sometido a la experiencia repetida de que tras sonar un silbato se le daba de comer. De esta manera el sonido del silbato por sí solo terminó por producir en el animal la respuesta de salivación asociada de manera natural a la comida.

El enfoque evolucionista del funcionalismo y el monismo materialista de la reflexología abrieron el camino a la aparición a comienzos del siglo XX del conductismo, un nuevo paradigma de la psicología de corte netamente positivista surgido como reacción contra el introspeccionismo y sus disputas interminables. Se suele considerar que el fundador del conductismo fue John Watson, psicólogo norteamericano que estudió en Harvard con William James y que desde el inicio de su carrera se interesó por la obra de Pavlov. Watson comenzó por redefinir el objeto de estudio de la psicología, afirmando que no debía ser la mente, sino sólo la conducta, dado que es el único fenómeno observable. La conducta es la respuesta del individuo a los estímulos que recibe del medio ambiente. Algunas conductas son simples y otras complejas, pero todas ellas han de ser explicables en función del esquema del arco reflejo. Esta tesis reduccionista, fundamental para los conductistas, resulta de aplicar el principio de parsimonia, según el cual no se deben multiplicar las entidades postuladas de manera innecesaria. El principio de parsimonia en psicología se conoce

como el *canon de Morgan*, que establece que «en ningún caso podemos interpretar una acción como el resultado de una facultad psíquica superior, si puede ser interpretada como el resultado de otra que se halla más abajo en la escala psicológica» (Carpintero, 1996, p. 279).

El propósito radicalmente positivista de Watson era convertir a la psicología en una ciencia de la naturaleza, sin conexión alguna con las ciencias sociales, y por tanto limitándose al método explicativo, al enfoque molecular, y a la causalidad eficiente. No es necesario apelar a la causalidad final debido a que las intenciones de los individuos, si es que las tienen, no son observables y por tanto no pueden ser objeto de estudio. Todas las conductas deben ser entendidas como respuestas a los estímulos del medio ambiente de una manera determinista. Dados los estímulos que recibe un sujeto, el psicólogo conductista podrá predecir su respuesta con la misma precisión con la que un astrónomo predice el movimiento de los planetas.

Como ya hemos visto, las respuestas pueden ser innatas o adquiridas, es decir, incondicionadas o condicionadas. Acerca de las respuestas condicionadas, Burrhus F. Skinner, el conductista más eminente después de Watson, realizó investigaciones de gran importancia. Basándose en el trabajo de Edward Thorndike sobre el aprendizaje asociativo, Skinner formuló la *ley del condicionamiento operante*. Mientras que en el condicionamiento clásico descubierto por Pavlov el estímulo precede a la conducta, en el condicionamiento operante sucede lo contrario: la conducta precede al estímulo. Se trata del procedimiento angular para el adiestramiento de animales. Cuando el animal produce espontáneamente una respuesta, se le proporciona una recompensa si se desea que la repita en el futuro, o bien un castigo para que no vuelva a hacerlo. De esta manera, los refuerzos permiten modelar la conducta de los individuos.

El paradigma conductista prevaleció en el ámbito académico anglosajón sobre sus competidores, tales como el psicoanálisis y la Gestalt, durante la primera mitad del siglo XX. A partir de la década de 1950 sus postulados teóricos comenzaron a debilitarse para dejar paso al cognitivismo, pero las aplicaciones prácticas de sus

descubrimientos siguen vigentes. La mercadotecnia, por ejemplo, se basa en buena medida en la ley del condicionamiento operante, pues las campañas de publicidad, además de ejercer un condicionamiento clásico, sirven para reforzar la conducta de comprar un determinado producto. Y al contrario, las conductas perjudiciales para los intereses de los poderosos son castigadas por éstos a través de las administraciones de Justicia con condenas severas convenientemente publicitadas para que el resto de la población tema imitarlas. Sin embargo, aunque el conductismo funciona bien con animales inferiores y con los seres humanos a cierto nivel, no es eficaz para dar cuenta de las conductas humanas superiores, y además excluye por principio el estudio de la vida mental. Por estas razones, hacia 1950 fueron varias las voces que se alzaron para proponer un nuevo paradigma explicativo de la psicología: el cognitivismo.

Supuestos nucleares

En 1948 se celebró en el Instituto de Tecnología de California un simposio multidisciplinar sobre los mecanismos cerebrales de la conducta que fue financiado por la Fundación Hixon, razón por la cual se lo conoce como el *Simposio de Hixon*. Allí, John von Neumann propuso la metáfora computacional, el matemático y neurofisiólogo Warren McCulloch expuso una teoría sobre cómo el cerebro procesa la información y el psicólogo Karl Lashley denunció la incapacidad del conductismo para explicar conductas organizadas complejas tales como jugar al tenis, tocar un instrumento musical y, sobre todo, expresarse en un lenguaje cualquiera (Gardner, 1985, p. 28). Estas ponencias sembraron las cuestiones necesarias para que en 1956 se celebrara un nuevo simposio en el cual, según el psicólogo George Miller, nació el paradigma cognitivista (Ibíd., p. 44). Éste tuvo lugar en el MIT, el Instituto Tecnológico de Massachusetts. En él participaron el propio Miller, el filósofo Noam Chomsky, y muchos otros, entre los que cabe destacar a los informáticos Allen Newell y Herbert Simon, recién llegados de haber celebrado un mes antes la conferencia de Dartmouth, el acontecimiento fundacional de la IA.

De todas las ideas propuestas durante aquellos primeros años en los que se gestó el cognitivismo nos interesan especialmente dos: la teoría de la información de Claude Shannon y Warren Weaver y el modelo de la neurona de Warren McCulloch y Walter Pitts. La primera afecta a la IA simbólica, mientras que la segunda es decisiva para los propósitos de la IA subsimbólica. La clave de la *teoría de la información* de Shannon y Weaver es que la información puede concebirse de forma totalmente independiente del contenido o la materia de que se trate como una decisión singular entre dos alternativas igualmente admisibles (Ibíd., p. 37). La unidad básica de información es el *bit*, acrónimo en inglés de dígito binario (*binary digit*). Un bit puede tener valor de 1 ó 0, como ya vimos al hablar de las máquinas de Turing. Dado un número *n* de bits, el número de alternativas posibles es igual a 2^n. En cuanto al *modelo de la neurona*, McCulloch y Pitts demostraron que las operaciones de una célula nerviosa y sus conexiones con otras formando redes de neuronas podían ser representadas mediante un modelo lógico (Ibíd., p. 34). Las neuronas se activan y a su vez activan a otras neuronas del mismo modo en que las proposiciones lógicas pueden implicar otras proposiciones. La intención de McCulloch y Pitts era, según sus propias palabras, «tratar el cerebro como una máquina de Turing» (Copeland, 2004, p. 408). Su modelo de la neurona fue decisivo para impulsar la confianza en la metáfora computacional de von Neumann.

Sobre el estudio de estas ideas y de muchas otras, Howard Gardner concluye que el paradigma cognitivista fue fundado sobre dos supuestos nucleares y otros tres que, si bien no son esenciales, son rasgos metodológicos que contribuyen a definirlo. Los dos supuestos nucleares ya los mencionamos al principio del presente capítulo. Se trata de la tesis internalista y de la metáfora computacional (García, 2001a, p. 18). La *tesis internalista* establece que para explicar adecuadamente la actividad humana o de cualquier otro sistema intencional que opere con causalidad final se requiere postular la existencia de procesos cognitivos, caracterizados como estados internos que dan cuenta, representan, conocen o informan de alguna realidad. Además, el cognitivismo reclama un nivel de análisis propio para estudiar dichos estados internos al margen de

los factores biológicos y de algunos de los factores ambientales que los afectan. Se trata de una tesis opuesta a la pretensión conductista de eliminar de la psicología el estudio de la vida psíquica. Por su lado, la *metáfora computacional*, también conocida como la *tesis del procesamiento de información*, caracteriza los procesos cognitivos como manipulación y operación con la información de modo análogo a como lo hace una computadora electrónica. En virtud de la teoría de la información de Shannon y Weaver se supone que el contenido de la información es irrelevante para su procesamiento. Daría igual, por tanto, que los estados internos del sistema intencional contuvieran representaciones simbólicas, analógicas, procedimentales o de cualquier otro tipo (García, 1996, p. 304).

Los otros tres rasgos del cognitivismo identificados por Gardner son el desprecio de algunos factores, la utilidad de los estudios interdisciplinarios y la continuidad temática, aunque no metodológica, con la tradición filosófica occidental (Gardner, 1985, p. 22). Comenzando por este último, Gardner se muestra como un firme defensor de la utilidad para los científicos cognitivistas que tiene el estudio de la tradición filosófica sobre el problema central de la epistemología y del cognitivismo: conocer el conocer. Sin embargo, se trata más de un deseo por parte de Gardner que de una realidad, ya que los científicos, como dice Steve Woolgar, están más preocupados por hacer que sus experimentos funcionen que de reflexionar sobre cuestiones epistemológicas, incluso en este caso en el que el objeto de estudio es el conocer mismo. La formación filosófica que los científicos reciben en la enseñanza reglada es escasa o nula. Raro es encontrar en la actualidad a científicos de gran talla como Newton, Einstein o el propio Gardner que dediquen parte de su tiempo a leer las reflexiones de los filósofos. Un filosofar de este tipo, dice Woolgar, «es más común entre los miembros más viejos y respetados de la comunidad o entre los desafectos y marginados de la misma» (Woolgar, 1988, p. 133).

Joseph Weizenbaum, ingeniero informático del MIT del que ya hemos hablado, se convirtió en un marginado en el momento en que publicó su libro *Computer power and human reason*, una obra en la que adoptaba un punto de vista crítico contra la

informática y la IA basándose en reflexiones tomadas de filósofos como Lewis Mumford (Weizenbaum, 1976, p. 29), Noam Chomsky (Ibíd., p. 118), Erich Fromm (Ibíd., p. 198) y Max Horkheimer (Ibíd., p. 205). La razón por la cual la filosofía suele ser rechazada en el seno de la ciencia es que la tarea del filósofo es la escenificada por Sócrates con su propia vida: ir de puerta en puerta haciendo ver a los demás que son más ignorantes de lo que se piensan. Y esto, al final, se paga como mínimo con la expulsión del grupo, como le ocurrió a Weizenbaum (Crevier, 1993, p. 143) y como pretendieron hacerle a Sócrates. La filosofía es una tarea demoledora de martillo y dinamita, como le gustaba decir a Nietzsche. En cambio, los científicos tienen un espíritu constructor. Desean construir su programa de investigación desarrollando la heurística positiva, al tiempo que defienden los supuestos del núcleo central ante cualquier ataque, tal y como les exige la heurística negativa. Son como artistas de castillos de arena que levantan barricadas o cinturones protectores para que el mar no destruya sus obras. De lo que no se dan cuenta es de que sus obras se secarán tarde o temprano, y entonces se desmoronarán. Cuando eso ocurra, necesitarán que el agua vuelva a mojar la tierra para poder construir algo nuevo con ella.

En cuanto a la utilidad de los estudios interdisciplinarios, el cognitivismo es un paradigma en el que convergen varias disciplinas. El nexo de unión de todas ellas es la psicología, que las aglutina gracias a su condición híbrida de ciencia biosocial que necesita de las aportaciones tanto de ciencias de la naturaleza como de ciencias sociales. Las seis ciencias cognitivas que Gardner destaca como las más importantes son: filosofía, lingüística, antropología, neurociencia, inteligencia artificial y, por supuesto, la propia psicología (Gardner, 1985, p. 53). El problema en este punto es el ya mencionado problema del método, y es que la psicología nunca ha podido conjugar los métodos explicativo y comprensivo de las ciencias naturales y sociales. La solución para los conductistas fue convertirla en una ciencia puramente natural. La solución del cognitivismo se supone, al menos nominalmente, que es la decisión salomónica de Wundt, que consiste en adoptar ambos puntos de vista, el natural y el social, al mismo tiempo. Sin embargo, la realidad es que los psicólogos cognitivistas están más cerca de

la solución eliminativista de los conductistas de lo que les gusta reconocer. La clave la encontramos en el tercer rasgo del cognitivismo que nos resta por reseñar: el desprecio de *algunos* factores. Decir "algunos" es un eufemismo.

Retorno al dualismo cartesiano

La psicología cognitiva reclama legitimidad para su pretensión de explicar la mente ignorando tres elementos, que son: la biología del cerebro en particular y del cuerpo en general; ciertos fenómenos de la vida psíquica como las emociones; y los factores ambientales más complejos de los que se ocupan las ciencias sociales (Ibíd., p. 22). Veámoslos en este mismo orden. Empezando por el cerebro, la metáfora computacional establece que se trata de un órgano equivalente al hardware de una computadora, mientras que la mente sería análoga al software. Los psicólogos cognitivistas argumentan que, al igual que los ingenieros informáticos desarrolladores de software no se ocupan de cuestiones de hardware, ellos pueden explicar la mente sin atender a su base biológica, formada por el cerebro y el cuerpo.

Atacando el problema de raíz, debemos denunciar que la tan manida distinción tajante ente hardware y software es falsa. Sobre este asunto Paul Ceruzzi dice: «La palabra "*software*" sugiere que hay una entidad singular, separada del *hardware* de la computadora, que trabaja con el *hardware* para solucionar un problema. La realidad es que no existe esa entidad singular. El sistema de una computadora es como una cebolla, con muchas capas distintas de *software* sobre un núcleo de *hardware*. Incluso en el centro –el nivel del procesador central– no hay una distinción clara: los *chips* que contienen microcódigo dirigen a otros *chips* para realizar las operaciones más básicas del procesador. Los ingenieros llaman a estos códigos "*firmware*", un término que sugiere lo borroso de la distinción» (Ceruzzi, 1998, p. 80). Éstas son las palabras de un historiador anglosajón especializado en informática y aeronáutica, alguien poco sospechoso, por tanto, de haber utilizado la analogía de la cebolla por influencia de un filósofo como Husserl.

También desde la neurofisiología encontramos voces críticas contra la separación de la mente y el cerebro defendida por la metáfora computacional. Es el caso de Antonio Damasio, quien denuncia que se trata de una versión contemporánea del dualismo de sustancias de Descartes. Sus defensores, dice Damasio, afirman que «la mente es el programa informático que se hace funcionar en un fragmento de equipo informático de ordenador llamado cerebro; o que cerebro y cuerpo están relacionados, pero sólo en el sentido de que el primero no puede sobrevivir sin el soporte vital del segundo» (Damasio, 1994, p. 284). Continúa señalando que: «Éste es el error de Descartes: la separación abismal entre el cuerpo y la mente. [...] Más específicamente: que las operaciones más refinadas están separadas de la estructura y funcionamiento de un organismo biológico» (Ibíd., p. 286).

Respecto a las *operaciones menos refinadas*, como por ejemplo el patrón rítmico de locomoción, los psicólogos cognitivistas no niegan las demostraciones científicas de que se trata de operaciones que dependen sólo de circuitos neuronales espinales, los cuales no necesitan información procedente del encéfalo. Si a una rata le seccionamos la médula espinal a la altura del cuello y la colocamos sobre una cinta andadora con la ayuda de un suspensorio para que mantenga el equilibrio, comenzará a caminar en cuanto la cinta andadora se mueva (Kandel, Schwartz & Jessell, 1995, p. 524). Pero en lo que se refiere a las *operaciones más refinadas*, como resolver una ecuación o recordar una imagen, los psicólogos cognitivistas reclaman que pueden ser explicadas a un nivel funcional independiente de la biología que las sustenta. Su planteamiento sería idéntico al de un comentarista de Fórmula 1 que no supiera nada de mecánica y que además alardease de su ignorancia, argumentando que no hace falta saber cómo funciona un motor de explosión para explicar lo que sucede en la pista. Si un vehículo se quedase sin gasolina, su explicación consistiría únicamente en decir que el coche se ha detenido por una razón desconocida e irrelevante.

Como dice el filósofo argentino Mario Bunge: «El autonomismo psicológico no sólo es científicamente esterilizante: también es impráctico, porque no puede ayudar a corregir los trastornos del comportamiento, de la afección o del aprendizaje. No puede

ser eficaz porque se empecina en imaginar que la mente es un ente separado del cuerpo, aunque admite que puede influir sobre éste. Este dualismo psicofísico le impide utilizar los recursos de la neurocirugía y de la psicofarmacología, así como las técnicas de modificación del comportamiento, ya que éstas se fundan sobre trabajos de laboratorio. ¡Pobre del maníaco-depresivo, del autista, del fóbico o del débil mental que caiga en manos de un logoterapeuta!» (Bunge, 1987, p. 23). Por nuestra parte, no defendemos que el análisis funcional de la vida psíquica deba ser despreciado, como propondría un conductista, sino que debe ser contemplado, pero siempre unido al análisis de los procesos biológicos subyacentes, tal y como hace la neuropsicología, que es el nuevo paradigma de la psicología que desde hace una década le está ganando terreno al cognitivismo. Cualquier otra postura a este respecto es un regreso al cartesianismo, y por tanto a una teoría de dualismo de sustancias que, además de ineficaz y hasta terapéuticamente peligrosa como denuncia Mario Bunge, no es científica porque no indica cómo puede ser falsada empíricamente.

En cuanto a los fenómenos mentales que el cognitivismo desprecia por principio, Gardner señala los afectos o emociones (Gardner, 1985, p. 22). Sobre este asunto John Haugeland apunta que los *estados de ánimo* (*moods*) son problemáticos para el cognitivismo porque no son en sí mismos representaciones mentales, ni simbólicas ni analógicas ni de ninguna clase (Haugeland, 1978, p. 271). La tristeza, por ejemplo, no está asociada a ninguna representación en particular. Sin embargo, un paradigma internalista como el cognitivismo no puede excluir de su examen a los estados de ánimo, porque se trata de fenómenos que afectan al modo en que se presentan las representaciones mentales. Así, en función del estado de ánimo una misma representación puede presentarse como una idea maravillosa o como una estupidez. Los cognitivistas son conscientes, por tanto, de que deben incluir las emociones en su explicación del funcionamiento de la mente, pero disminuyen su importancia argumentando que la ciencia cognitiva sería impracticable si se quisiera tomar en cuenta todos estos elementos fenoménicos individualizadores: «Al querer explicarlo todo, se termina por no explicar nada» (Gardner, 1985, p. 58).

El centauro de Platón

Bajo esta manera de pensar se esconde la presunción racionalista de que la razón y las pasiones pueden funcionar de manera independiente. Esto, a día de hoy y con demostraciones empíricas, se sabe que es falso. En vez de pensar, como hacía Platón, en la razón y las pasiones como un auriga y dos caballos, el uno blanco y el otro negro, la metáfora que más se ajusta a la realidad es la de un centauro, teniendo el auriga un cuerpo de caballo con manchas blancas y negras. A lo largo de la evolución de las especies el auriga se desarrolló sobre el cuerpo del caballo, y no al revés, pues las pasiones son evolutivamente anteriores a la razón, y la evolución procede utilizando las estructuras más antiguas para desarrollar sobre ellas las posteriores. Basándose en este hecho, evidenciado a mediados del siglo XX por las investigaciones del *cerebro trino* del neurocientífico Paul MacLean, Damasio advierte que «el aparato de la racionalidad, que tradicionalmente se suponía que era *neo*cortical, no parece funcionar sin el de la regulación biológica, que tradicionalmente se presumía que era *sub*cortical. La naturaleza parece haber construido el aparato de la racionalidad no sólo encima del aparato de la regulación biológica, sino también *a partir* de éste y *con* éste» (Damasio, 1994, p. 155).

Por tanto, la racionalidad verdadera en toda su amplitud, que es la que está presente en cada acto de la vida, necesita de las pasiones. Prueba de ello es el célebre caso de Phineas Gage, un obrero del ferrocarril de mediados del siglo XIX que, por culpa de un accidente, sufrió una lesión cerebral terrible. Resumiéndolo mucho, la lesión desconectó las áreas subcorticales responsables de las emociones respecto de las áreas neocorticales del pensamiento racional. La consecuencia fue que Gage quedó incapacitado para comportarse de manera racional. Podía resolver problemas matemáticos, caminar, hablar, y mostrarse como un ser humano en apariencia normal a los ojos de un médico que lo diagnosticase durante sólo unos minutos. Pero en una escala de tiempo mayor se observaba que Gage había perdido muchas facultades

importantes: «La capacidad de anticipar el futuro y de planear en consecuencia dentro de un ambiente social complejo; el sentido de la responsabilidad hacia uno mismo y hacia los demás; y la capacidad de orquestar deliberadamente la propia supervivencia, y el control del libre albedrío de uno mismo» (Ibíd., p. 29). Incapaz de comportarse como un ser humano en sociedad, Gage perdió su empleo en el ferrocarril y sólo pudo ganarse la vida como monstruo circense, deambulando de aquí para allá, hasta que la muerte acabó con su desdichada existencia.

En términos funcionales, lo que el cerebro de Gage no tenía en su proceso de toma de decisiones era la función del *marcador somático* (*somatic marker*). El sistema neuronal para la adquisición de señales de marcadores somáticos se halla en las cortezas prefrontales, justo las que Gage tenía gravemente dañadas. Gracias a la función del marcador somático, cuando nos enfrentamos a un problema, antes de razonar conscientemente hacia su solución, las opciones son evaluadas a nivel inconsciente como mejores o peores en función de los sentimientos asociados a los resultados a los que condujeron en ocasiones anteriores (Ibíd., p. 205). El marcador somático es imprescindible para la toma de decisiones debido a que la atención y la memoria funcional, que son las estructuras encargadas del cálculo puramente racional, tienen una capacidad limitada. Es lógico que los cognitivistas pretendan afrontar las dificultades aisladamente, de manera cartesiana. Pero la realidad es que una psicología como la suya, que disminuye la importancia de las emociones hasta el punto de muchas veces ignorarlas, está condenada a producir modelos de la mente de individuos deficientes y perturbados como Phineas Gage. Sería más razonable imitar el proceso de la evolución y comenzar estudiando lo anterior, las emociones, para construir sobre su conocimiento una teoría del raciocinio.

Finalmente, el tercer elemento que los cognitivistas desprecian en su estudio de la mente es el de los factores históricos y culturales, y el papel del contexto o de los antecedentes (Gardner, 1985, p. 22). Es llamativo que Gardner justifique esta actitud en líneas generales al tiempo que carga contra las teorías que ignoran el carácter contextual de la inteligencia. De su colega Hans Eysenck llega a decir que su pretensión

de medir la inteligencia con un casco de electrodos es grotesca (Gardner, 1993, pp. 279 y 289). Atrincherado en su bastión fisicalista, desde el que ha conseguido convertirse en uno de los psicólogos más prestigiosos del Reino Unido, Eysenck afirma que la única manera científica de estudiar la inteligencia pasa por eliminar los factores más complejos, tales como la educación, la motivación, la posición socio-económica, la personalidad, la influencia familiar, la nutrición y la motivación (Eysenck, 1986, p. 69). La pretensión similar de los cognitivistas de excluir los factores ambientales complejos se debe al ya mencionado problema del método. Ante la dificultad, o tal vez incluso la imposibilidad, de conjugar en una sola explicación integral los factores naturales y sociales que modelan la mente, las escuelas de psicología tienden a adoptar un enfoque eliminativista, y se sitúan en uno de los dos polos.

Los conductistas son un caso proverbial de la intención positivista de convertir a la psicología en una ciencia puramente natural, mientras que en el otro extremo encontramos la tradición humanista y filosófica representada por Dilthey, Brentano, Husserl y Heidegger, entre otros. La primera corriente, de método explicativo y enfoque molecular, explica eficazmente los mecanismos más simples de la vida psíquica, tales como el condicionamiento clásico y el condicionamiento operante, pero no alcanza a las funciones mentales superiores como el lenguaje. Por el otro lado, el método comprensivo, de enfoque molar, describe bien las funciones mentales superiores, pero no es capaz de explicar cómo surgen a partir de elementos más simples, sino que se queda en la mera descripción. Sobre el método comprensivo Gardner reconoce que «cabe preguntarse si esta rama de la psicología ha logrado algo que no fuera notorio para nuestros antecesores, o aun para cualquier observador lego en psicología. Después de todo, no es menester la tecnología del siglo XX ni elaboradas estadísticas para demostrar que los individuos aportan a cada nueva actividad sus experiencias y sus marcos organizadores, o que es posible procesar la información con mayor o menor empeño según los diversos propósitos» (Gardner, 1985, p. 146). La psicología ha estado siempre atrapada en el dilema de elegir entre uno de los dos enfoques, y el paradigma cognitivista no es una excepción.

Pretensiones positivistas

Como vemos, la solución de la psicología cognitiva consiste en huir de lo más problemático: ignora la biología del cerebro y del cuerpo, disminuye la importancia de ciertos fenómenos mentales como las emociones y se abstrae de las condiciones ambientales complejas. La razón por la cual la psicología cognitiva se ve obligada a ignorar algunos factores reside en que su objetivo, como el de toda ciencia, es elaborar *teorías* o *modelos*, y estos conceptos implican siempre una simplificación. La distinción entre teoría y modelo es problemática. A veces se ha sugerido que hay diferencias entre ambos conceptos, y a veces se ha dicho que un modelo es equivalente a una teoría (Ferrater, 1965c, p. 216). Otros puntos de vista sostienen que el modelo es un constructo intermedio bidireccional entre los datos y la teoría (García, 2001a, p. 78). En cualquier caso, los modelos tienen tres rasgos definitorios que, a nuestro juicio, comparten con las teorías. En primer lugar, son una duplicación de la realidad modelada con la que guarda algún tipo de correspondencia, analógica o simbólica. Segundo, son una simplificación de esa realidad seleccionando o sobrevalorando algunos rasgos con el consiguiente riesgo de simplificación o reduccionismo. Y tercero, tienen un carácter sistémico más o menos explícito de los distintos niveles de análisis, estructural, procesual y funcional (Ibíd., p. 77).

La psicología cognitiva justifica su desprecio de los factores que hemos señalado en virtud de la segunda característica de los modelos y las teorías, y es que se trata de simplificaciones. Así, ya mencionamos en el capítulo anterior que en las ecuaciones para calcular un tiro parabólico se ignora el color del proyectil. En realidad es una variable que sí afecta a su velocidad, pero lo hace en un orden de magnitud tan pequeño que se desprecia, así como a veces se desprecia el rozamiento del aire. Como dice Gardner, la psicología cognitiva necesita restringir los factores que afectan a la mente para poder elaborar modelos y teorías sobre ella, pues si se quisiera explicarlo todo, se terminaría por no explicar nada (Gardner, 1985, p. 58).

La cuestión es el grado de relevancia de los factores que se desprecian. El color afecta muy poco a un proyectil, pero es innegable que la biología del cerebro, las emociones y las influencias culturales afectan mucho a la mente. Si los psicólogos cognitivistas realizaran la selección de factores en función de criterios teóricos, es obvio que no podrían dejar fuera a estos tres. Sin embargo, como indica el psicólogo especialista en IA Zenon Pylyshyn, la validación de modelos implica siempre asunciones metateóricas (Pylyshyn, 1974, p. 93). La asunción metateórica que conduce a los psicólogos cognitivistas a despreciar los factores señalados es la metáfora computacional. En realidad no hay ninguna razón teórica para creer que la mente funciona como un procesador de información similar a una computadora electrónica. Se trata sólo de un supuesto adoptado con el propósito de convertir a la psicología en una ciencia tan exitosa como las ciencias de la naturaleza. Si la mente procesara la información como una computadora electrónica, eso implicaría que las leyes del pensamiento serían expresables en un lenguaje de programación. Y los lenguajes de programación, como vimos en el capítulo anterior, son lenguajes formales. De esta manera, la metáfora computacional promete a los psicólogos que se adhieran a ella una capacidad explicativa de la mente tan exacta como la que anhelaban los conductistas pero sin renunciar a la tesis internalista. Los fenómenos que no se dejan formalizar son despreciados. Ésta es la verdadera razón por la cual el cognitivismo desprecia las emociones y las influencias culturales: porque, al no dejarse formalizar, se interponen en el camino hacia una psicología positivista que aspira a formular sus explicaciones en el estilo nomológico-deductivo propio de las ciencias de la naturaleza más duras, como la física y la química (Haugeland, 1978, p. 244).

Gardner reconoce que en antropología han fracasado los intentos por formalizar las influencias culturales. Pone como ejemplo el caso de un estudio etnográfico llevado a cabo en Chiapas, Méjico, sobre el consumo de bebidas alcohólicas. El objetivo era analizar las pautas vinculadas al hábito de la bebida. Uno de los miembros del equipo investigador, Paul Kay, dejó escrito el testimonio siguiente: «Ocurrió que tras reunir una enorme cantidad de material y de pasar dos o tres años

buscando una serie de procedimientos objetivos (o al menos semi objetivos), que permitieran reducirlos a alguna suerte de enunciación lógica, debimos renunciar, porque no pudimos hallar esa serie de procedimientos. En Chiapas, la bebida es una institución que impregna toda la vida de la gente: la religión, la política, la vida familiar y hasta la agricultura están inextricablemente ligadas a ella; de modo tal que hacer la etnografía de la ingestión de bebidas alcohólicas equivale a hacer la etnografía total de este pueblo» (Gardner, 1985, p. 277). Por seguir con el símil del tiro parabólico, los investigadores de Chiapas se encontraron con que a su particular tiro parabólico le afectaba todo de manera relevante: el color del proyectil, el color del cielo, el color de la pólvora, la marca del reloj del hombre que encendía la mecha, el olor de las flores del campo de tiro, el número de árboles de hoja caduca en las inmediaciones, y así hasta el infinito. Evidentemente, no pudieron formalizar el fenómeno. La teoría de la información de Shannon y Weaver postula que toda información, incluida por tanto la cultural, es codificable en un lenguaje formal de alfabeto binario, pero Paul Kay y sus colegas descubrieron que la realidad, como decía Lenin, es tozuda.

En la misma línea, dentro de la lingüística, que es otra de las disciplinas que colaboran con la psicología dentro del paradigma cognitivista, Chomsky presentó en el Simposio de 1956 en el MIT una teoría del lenguaje que pretendía estudiar la sintaxis de manera independiente a la semántica y la pragmática (Ibíd., p. 206). No nos gustaría que se entendiera esta crítica como una diatriba contra la obra de pensadores tan geniales como Chomsky. Nuestra intención es evidenciar el prejuicio operante en los cimientos del paradigma cognitivista: que la mente es una computadora electrónica para así convertir a la psicología en una ciencia formal. Gardner lo reconoce abiertamente a propósito de la antropología: «El éxito del enfoque cognitivo en antropología dependerá de que el rigor del análisis de componentes, o de cualquier otro método de inspiración computacional, pueda amalgamarse con el abordaje de los problemas generales que tradicionalmente interesaron a los estudiosos de culturas extrañas» (Ibíd., p. 282). Donde dice "amalgamarse" se está refiriendo al problema del método. Ante la imposibilidad de conjugar los métodos explicativo y comprensivo, los

psicólogos cognitivistas utilizan la metáfora computacional como excusa para acercarse a la simplificadora univocidad de los enunciados las ciencias de la naturaleza y alejarse de la enrevesada multivocidad de los textos de las ciencias sociales.

En definitiva, el objeto de estudio que resulta de despreciar la importancia del cerebro en particular y del cuerpo en general, de ignorar ciertos fenómenos mentales como las emociones, y de subestimar las complejidades culturales, es una mente escuálida, reducida a la condición de un sistema formal como lo es cualquier programa informático. El programa de investigación de la IA simbólica, debido a su incardinación original en el paradigma cognitivista, pretende utilizar computadoras electrónicas para convertirlas en mentes artificiales semejantes a computadoras electrónicas. Ésta es una circularidad encubierta que ha resultado improductiva durante décadas, como veremos en el capítulo séptimo. Por ahora, en la siguiente sección vamos a examinar las ideas centrales de la neurociencia, que es la disciplina en la que se basa el otro gran programa de investigación de la inteligencia artificial: la IA subsimbólica, que aspira a construir máquinas inteligentes reproduciendo o simulando el funcionamiento de las redes de neuronas del cerebro.

4.3. Neurociencia

La *neurociencia* estudia la estructura y organización funcional del sistema nervioso (García, 2001a, p. 24), o lo que es lo mismo, estudia el sistema nervioso (SN) a nivel estático y dinámico. El sistema nervioso está compuesto de dos tipos de células: neuronas y células gliales. Las *neuronas* transportan información en forma de señales eléctricas, mientras que las *células gliales*, también llamadas *glia*, realizan funciones secundarias al servicio de las neuronas. En las exposiciones del sistema nervioso suele tratarse en primer lugar la anatomía y funcionamiento de las neuronas, pero aquí vamos a comenzar por el sistema nervioso dando por supuesto que todo el mundo sabe, a grandes rasgos, qué son las neuronas y cómo funcionan. Las neuronas al detalle las veremos después, en una sección aparte.

El *sistema nervioso* se divide en el sistema nervioso central (SNC), compuesto por el cerebro y la médula espinal, y el sistema nervioso periférico (SNP), que comprende los ganglios y los nervios periféricos que están fuera del cerebro y de la médula espinal (Kandel, Schwartz & Jessell, 1995, p. 77). A su vez, el *sistema nervioso periférico* tiene una división autónoma y otra somática. La *división somática* proporciona al sistema nervioso central la información sensorial interoceptiva y exteroceptiva al tiempo que transporta la información motora en sentido opuesto. La *división autónoma* se conoce también con el nombre de *sistema motor autónomo*, y consiste en tres partes diferenciadas físicamente, que son los sistemas nerviosos simpático, parasimpático y entérico (Ibíd., p. 598). La función del *sistema simpático* se conoce en inglés con la rima mnemotécnica *fight or flight*, que indica que es el encargado de activar los recursos del organismo ante situaciones de emergencia, tales como luchar o salir volando para escapar de un peligro. La función del *sistema parasimpático* es la expresada en otra rima, la que dice *rest and digest*, y que deja claro que es el sistema que se ocupa de administrar los recursos para evitar su agotamiento y restablecer el equilibrio en los estados de reposo. Por último, el *sistema nervioso entérico* controla la función del músculo liso del tubo digestivo.

El *sistema nervioso central* está organizado en tres ejes que sirven para describir la posición de sus componentes. Suponiendo que estuviéramos observando a un cuadrúpedo, el *eje rostro-caudal* es el que va del rostro hacia la cola, el *eje dorso-ventral* es el que va del lomo al vientre, y el *eje latero-medial* es el que va desde el punto medio del cuerpo a cualquiera de los lados, izquierda o derecha. En el caso del ser humano, debido a su posición bípeda, el eje rostro caudal no es completamente recto. Asciende en vertical desde el final de la espalda hasta el diencéfalo, que está aproximadamente a la altura de los oídos, y allí se curva en un ángulo de noventa grados para tomar la dirección horizontal que va hacia el rostro. El sistema nervioso central se divide en siete regiones anatómicas que, enumeradas en un orden aproximadamente caudo-rostral, son: médula espinal, bulbo raquídeo, protuberancia, cerebelo, mesencéfalo, diencéfalo y hemisferios cerebrales (Ibíd., p. 77).

La *médula espinal* tiene una vía *ascendente* o *aferente* por la cual transporta la información sensorial desde el sistema nervioso periférico hasta las partes superiores del sistema nervioso central, mientras que la vía *descendente* o *eferente* transporta la información motora en sentido inverso, es decir, del sistema nervioso central al sistema nervioso periférico. La médula espinal se compone de treinta y un pares de nervios espinales, que son nervios periféricos formados por la unión de raíces dorsales y ventrales. Las dorsales conducen la información sensorial, mientras que las ventrales comunican la información motora.

Encima de la médula espinal se encuentra el *tallo cerebral*, que se compone del bulbo raquídeo, la protuberancia y el mesencéfalo. El *bulbo raquídeo* regula funciones autónomas imprescindibles para la vida, tales como la presión sanguínea y la respiración. La *protuberancia* es una vía de comunicación directa de los hemisferios cerebrales con el cerebelo. En posición dorsal respecto de la protuberancia se encuentra el *cerebelo*, una estructura muy compleja dividida en lóbulos que recibe información sensorial de la médula espinal, información motora de la corteza cerebral, e información del sistema vestibular acerca del equilibrio del cuerpo. Con todas esas entradas, el cerebelo actúa como un comparador del estado actual del cuerpo con la información motora saliente, ajustando esta última para afinar la coordinación corporal. En consecuencia, las lesiones del cerebelo no son paralizantes, pero dificultan el equilibrio y la realización de movimientos precisos y coordinados (Ibíd., p. 536). El *mesencéfalo* está en posición rostral a la protuberancia, y es la parte más pequeña de las tres que componen el tallo cerebral. Está implicado en el movimiento de los ojos y de algunos músculos, además de ser un centro de relevo de las señales auditivas y visuales hacia los hemisferios cerebrales. Sobre el mesencéfalo descansa el *diencéfalo*, que se divide en el tálamo y el hipotálamo. El *tálamo* es el centro de relevo más importante del sistema nervioso. Procesa y distribuye casi toda la información sensorial y motora que va a la corteza cerebral. Por su parte, el *hipotálamo*, situado en posición ventral respecto del tálamo, regula el sistema nervioso autónomo y la secreción de hormonas mediante la glándula pituitaria (Ibíd., p. 97).

La séptima y más grande de todas las partes anatómicas del sistema nervioso central son los *hemisferios cerebrales*, que se dividen en la corteza cerebral por un lado, y tres estructuras profundas por el otro, que son los ganglios basales, la formación hipocampal y la amígdala. Los *ganglios basales* consisten en cinco núcleos densamente interconectados: núcleo caudado, putamen, globo pálido, núcleo subtalámico y sustancia negra. Los ganglios basales están muy relacionados con la función del cerebelo, ya que también influyen sobre el movimiento del tronco y las extremidades. Por ejemplo, la enfermedad de Parkinson, caracterizada por los temblores, se produce por una degeneración de las vías dopaminérgicas que la sustancia negra proyecta sobre el estriado, que es la estructura compuesta por el núcleo caudado y el putamen (Ibíd., p. 547).

La *formación hipocampal* interviene en la formación de la *memoria explícita o declarativa* a largo plazo (Ibíd., p. 680), que es la memoria que contiene información acerca del mundo, bien en forma de *memorias episódicas* acerca de sucesos particulares o bien en forma de *memorias semánticas* sobre generalizaciones, frente a la *memoria implícita o procedimental*, que normalmente no puede expresarse en palabras, y que puede ser no-asociativa, como la habituación y la desensibilización, o bien asociativa, como el condicionamiento clásico y el condicionamiento operante. La formación hipocampal está también implicada en las emociones junto con la tercera estructura profunda, la amígdala. Ambas forman parte del *sistema límbico*, que es un concepto auxiliar bastante difuso en el que se incluyen aquellas partes del sistema nervioso que intervienen en las emociones.

De todas esas partes, la *amígdala* es la más específicamente relacionada con el ámbito de lo emocional (Ibíd., p. 608). Se divide en varios núcleos que están recíprocamente conectados con el hipotálamo, la formación hipocampal, el neocórtex y el tálamo. El núcleo central de la amígdala proyecta sobre las áreas corticales de asociación, especialmente la corteza orbitofrontal y el giro cingulado, siendo estas proyecciones unas vías fundamentales para la percepción consciente de las emociones. A nivel neurofisiológico la lesión de Phineas Gage en el lóbulo frontal del

cerebro resultó en una incapacidad para asociar emociones a determinadas categorías de situaciones y estímulos (Damasio, 1994, p. 166), a causa de lo cual era incapaz de comportarse racionalmente. Una lesión no ya en las vías de comunicación de la amígdala, sino en la amígdala misma, tiene efectos aún más devastadores, pues se deteriora el procesamiento de las emociones primarias, que son las innatas o incondicionadas, como la que produce la salivación del perro al ver la comida.

La corteza cerebral

La *corteza cerebral* es la estructura evolutivamente más reciente del sistema nervioso central, y es en ella donde suele ubicarse la inteligencia (Hawkins & Blakeslee, 2004, p. 55). Sin embargo, a la luz del caso de Phineas Gage y del examen de las teorías contemporáneas de la inteligencia que realizaremos en el próximo capítulo, se torna evidente que identificar la corteza como la única sede de la inteligencia es un error. El centauro tiene una parte racional en forma de auriga y otra pasional en forma de caballo. Separando ambas por la mitad lo que resulta no es ni un auriga completo ni un caballo completo, sino un solo cadáver partido en dos.

La corteza está compuesta por alrededor de treinta mil millones de neuronas (Ibíd., p. 58), $3x10^{10}$, lo cual supone casi la tercera parte de las 10^{11} que hay en el cerebro (Kandel, Schwartz & Jessell, 1995, p. 21). Su superficie es de 1,5 m^2 y tiene un grosor medio de dos milímetros en los que se distinguen seis capas. Para maximizar la extensión de superficie cortical dentro del cráneo sin que éste tuviera que aumentar su volumen, la evolución adoptó la estrategia de plegar la corteza sobre sí misma, formando *giros* o *circunvalaciones*, que son las zonas en relieve, y *surcos*, que son las hendiduras. Los surcos más grandes y pronunciados dividen la corteza de cada hemisferio en cuatro *lóbulos* que toman sus nombres de los huesos craneales bajo los que se descansan: frontal, parietal, temporal y occipital. El *frontal* es que se encuentra sobre los ojos, el *occipital* está en la parte posterior de la cabeza, el *parietal* está entre ambos en posición dorsal, y el *temporal* está entre ambos en posición ventral.

A principios del siglo XX, el anatomista alemán Korbinian Brodmann dividió la corteza en 52 áreas atendiendo a un criterio citoarquitectónico, es decir, a la forma de las neuronas predominantes en cada lugar. Aunque posteriormente se ha demostrado que esta división no es funcionalmente precisa, se sigue utilizando como topografía preferente para distinguir las áreas de la corteza. Cada una de las áreas de Brodmann está asociada a una o varias funciones, siendo todas ellas de carácter elemental. Las funciones que más extensión de corteza ocupan son las de procesamiento de la información sensorial y motora (Ibíd., p. 83). En función de la complejidad creciente de esa tarea de procesamiento, las áreas corticales se diferencian en primarias, secundarias y terciarias. La cúspide del proceso reside en las *áreas de asociación*, que integran la información sensorial y motora de diversos tipos con la información procedente del sistema motivacional para crear las representaciones mentales con las que se realizan todos los actos intencionales, ya sean sencillos o complejos.

Las áreas de asociación más importantes son la corteza parieto-témporo-occipital, la corteza asociativa prefrontal y la corteza asociativa límbica. La *corteza parieto-témporo-occipital* está situada en la intersección de los lóbulos indicados por su nombre. Se encarga de las funciones perceptuales de alto nivel relacionadas con el tacto, el oído y la visión, que son sentidos cuyas áreas de procesamiento primario residen respectivamente en los lóbulos parietales, temporales y occipitales. La *corteza de asociación prefrontal* ocupa la mayor extensión de la parte rostral del lóbulo frontal, y es donde se realiza la planificación del movimiento voluntario. Y, por último, la *corteza de asociación límbica* abarca regiones de los lóbulos parietales, frontales y temporales, y está implicada en la motivación, la emoción y la memoria.

Principios de organización funcional

El sistema nervioso central, además de dividirse en las siete partes que acabamos de exponer, también se divide en tres grandes sistemas funcionales: sensorial, motor y motivacional. Todos ellos cumplen al menos con seis principios de

organización funcional, que son el relevo sináptico, el procesamiento en paralelo, la organización topográfica, la decusación, las dimensiones funcionales y la organización modular (García, 2001a, p. 263). Los *núcleos de relevo* son centros de procesamiento donde la información entrante es modificada por la interacción entre *interneuronas locales* y enviada posteriormente a otros lugares mediante *interneuronas de proyección* o *principales*. Hay centros de relevo distribuidos por todo el sistema nervioso central, siendo el tálamo el más importante, dado que es el que procesa casi toda la información sensorial que llega a la corteza, con la única excepción de la información olfativa, que llega directamente debido a que el bulbo olfatorio es en sí mismo una forma primitiva de corteza (Kandel, Schwartz & Jessell, 1995, p., 384). El olfato es, pues, como decía Nietzsche, un sentido muy especial.

En cuanto al *procesamiento en paralelo*, los sistemas sensorial, motor y motivacional están compuestos de subsistemas que son independientes a nivel anatómico y funcional. Por ejemplo, el tacto, que técnicamente se conoce con el nombre de sistema somatosensorial, es una modalidad sensorial con cuatro submodalidades, que son el tacto, la propiocepción o percepción de la posición de los miembros, el dolor y la temperatura. Cada una es captada por receptores específicos y conducida por vías paralelas diferentes hacia el sistema nervioso central, donde también se procesan de manera separada (Ibíd., p. 375).

Respecto a la *organización topográfica*, sólo se cumple en las modalidades somatosensorial, visual y auditiva. En ellas la organización espacial de los receptores sensoriales es preservada a lo largo del recorrido de la información por el sistema nervioso central. En el caso de la visión, los fotorreceptores de la retina forman un imagen bidimensional que se conserva en todo momento hasta que la información llega a las cortezas inferotemporales, que es donde se produce el reconocimiento de las imágenes (Ibíd., p. 445). Gracias a este fenómeno, con una máquina de resonancia magnética funcional (fMRI) se puede ver aproximadamente las mismas imágenes que otro ser humano está viendo o rememorando mediante la monitorización de su corteza visual primaria, situada en la región más caudal del lóbulo occipital. El principio

de organización topográfica también se cumple en el sistema motor. En la corteza motora primaria, que está situada en la circunvalación anterior al gran surco central que separa los lóbulos frontal y parietal, se encuentran representados todos los músculos esqueléticos del cuerpo humano, que son los que se pueden accionar voluntariamente. Y justo detrás, en la circunvalación posterior, está la corteza somatosensorial primaria, con una representación de todos los receptores sensoriales.

El cuarto principio de organización funcional del sistema nervioso central que hemos enumerado es el de *decusación*, que se refiere al hecho de que muchas vías neuronales cambian de lado en algún momento de su recorrido. Esos lugares de cruce contralateral son las decusaciones, y las estructuras formadas por la convergencia de varias decusaciones se denominan comisuras. Debido al fenómeno de la decusación, las sensaciones y los movimientos del lado izquierdo del cuerpo son procesados por el hemisferio derecho, y los del lado derecho, por el hemisferio izquierdo. Una excepción a esta regla es el sistema visual, ya que la retina de cada ojo se divide verticalmente en dos hemirretinas, la nasal, que es la que está más cerca de la nariz, y la temporal, que es la otra. En cada ojo su hemirretina nasal proyecta contralateralmente a través de la decusación del quiasma óptico, mientras que la hemirretina temporal proyecta ipsilateralmente, es decir, al hemisferio del mismo lado en el que se encuentra ese ojo (Ibíd., p. 427).

Respecto a las *dimensiones funcionales*, las funciones del cerebro pueden ser divididas en tres dimensiones: superior-inferior, anterior-posterior e izquierdo-derecho (García, 2001a, p. 266). Hay que tener en cuenta que ésta es una división puramente analítica, dado que en la realidad el cerebro trabaja como una totalidad cuya división quirúrgica produce, como mínimo, graves trastornos. La parte superior alberga los procesos psíquicos superiores, y la inferior, el resto. Las otras dos dimensiones se aplican sólo al cerebro superior. La anterior se ocupa del movimiento y la planificación, mientras que la posterior está más relacionada con la percepción. Y finalmente, en el hemisferio izquierdo predominan los procesos secuenciales, como el lenguaje, y en el derecho los procesos en paralelo, como la representación espacial.

En cuanto a la *organización modular*, que es el último de los seis principios organizativos que hemos señalado, se refiere al hecho de que el cerebro está organizado en múltiples subsistemas, redes, circuitos neuronales y módulos con un funcionamiento relativamente independiente y actuando en paralelo (Ibíd., p. 263). De esta manera, un sujeto puede sufrir una lesión en el lóbulo occipital que perjudique a su percepción visual, que es la facultad localizada en esa región, pero sin que afecte, por ejemplo, a su competencia lingüística. Hoy en día disponemos de numerosas evidencias experimentales en favor de la modularidad del cerebro, pero en épocas pasadas algunos investigadores creyeron que era un órgano homogéneo en el que la totalidad participaba en todas las funciones.

Holismo y localizacionismo

En la primera mitad del siglo XIX había dos concepciones diametralmente opuestas acerca de la localización de las funciones en el cerebro. Por un lado, el neuroanatomista alemán Joseph Gall sostenía que las funciones superiores más abstractas y complejas estaban localizadas en la corteza. Así, por ejemplo, creía que la firmeza del carácter dependía de una zona ubicada en el lóbulo parietal, y que la noción de causalidad era producida por una pequeña región del lóbulo frontal. Gall estaba convencido de que la prevalencia de un rasgo de personalidad se reflejaba en un mayor volumen cortical de su área correspondiente, lo cual generaba a su vez una deformación ostensible del cráneo. De esta manera, según la doctrina frenológica que él impartía, se podía determinar la personalidad de un sujeto midiendo su cráneo.

Contra este localizacionismo extremo de las funciones mentales más refinadas, el médico Pierre Flourens propuso en la década de 1830 la *teoría del campo agregado*, según la cual todas las regiones de la corteza participan en todas las funciones mentales. Así, la gravedad de las deficiencias mentales de un sujeto dependería no del área cortical en la que se hubiese producido la lesión, sino de la cantidad total de tejido cortical lesionado. Poco después, en la segunda mitad del XIX, el cirujano francés

Paul Broca y el neurólogo alemán Carl Wernicke realizaron sendos hallazgos por separado que, si bien no apoyaban un localizacionismo tan extremo como el de Gall, refutaban la teoría del campo agregado, en tanto que demostraban que ciertas funciones lingüísticas están localizadas en la corteza. En la misma línea, el neurólogo británico John Hughlings Jackson realizó importantes investigaciones sobre la epilepsia focal. No obstante, los esfuerzos de Hughlings Jackson no bastaron para convencer a sus contemporáneos, y además los estudios de Broca y Wernicke contenían algunas deficiencias que fueron utilizadas por los seguidores de Flourens para seguir manteniendo la teoría del campo agregado como tesis fundamental de la neurociencia hasta principios del siglo XX.

En esa época, Charles Sherrington y Santiago Ramón y Cajal hicieron resurgir el localizacionismo moderado con la *teoría del conexionismo celular*, la cual establece que las neuronas son las unidades de señalización del cerebro y están dispuestas generalmente en grupos funcionales que se conectan de forma precisa (Kandel, Schwartz & Jessell, 1995, p. 8). Hubo investigadores, como Karl Lashley, que continuaron defendiendo el holismo del campo agregado todavía en el siglo XX, pero finalmente tuvieron que rendirse a la evidencia de los hallazgos localizacionistas, como por ejemplo los de Roger Sperry sobre pacientes comisurotomizados, es decir, con los hemisferios cerebrales desconectados entre sí.

En la actualidad nadie discute ya que la corteza está dividida en áreas que realizan funciones diferentes, pero al mismo tiempo se reconoce que la localización es sólo de funciones muy elementales y que la corteza, tal y como creían los partidarios de la teoría del campo agregado, posee una innegable plasticidad. Un caso muy claro de plasticidad lo encontramos en los sujetos que se quedan ciegos en la edad adulta. Antes de adquirir la minusvalía, su lóbulo occipital se ocupaba del procesamiento de la información visual. Al quedarse ciegos por una enfermedad de los ojos y no haber ya más información visual entrante, ese mismo área pasa a dedicarse, tras un cierto tiempo de adaptación, al procesamiento de información táctil para leer en braille. Ciertamente, los niños y las mujeres son las poblaciones con mayor plasticidad, pero

los adultos y los varones también poseen algún grado hasta el final de su vida. La plasticidad es un hecho recogido por muchas teorías, entre las que destacan el anteriormente citado principio de Hebb y la hipótesis de Mountcastle.

El *principio de Hebb*, también conocido como el *principio de la convergencia sincrónica* (Fuster, 1997, p. 74), establece que cuando una neurona A participa repetida o persistentemente en la excitación o inhibición de otra neurona B, entonces acontece algún tipo de proceso o cambio metabólico en una o ambas que incrementa la eficacia de A para excitar o inhibir a B (Kandel, Schwartz & Jessell, 1995, p. 681). Y a la inversa, si la activación sincrónica de varias neuronas remite, entonces dejan de formar una red común, que es lo que sucede cuando dejamos de rememorar un recuerdo y lo olvidamos. En inglés hay un juego de palabras que lo resume en una escueta frase: *neurons that fire together, wire together* (las neuronas que disparan juntas, se conectan juntas). Por su parte, la hipótesis formulada por Vernon Mountcastle sostiene que la corteza cerebral usa la misma herramienta computacional para realizar todo lo que hace (Hawkins & Blakeslee, 2004, p. 66). En virtud de lo observado en el fenómeno de la plasticidad neuronal, la hipótesis de Mountcastle tiene bastante crédito. Volveremos a esta hipótesis en el capítulo quinto, a propósito de la teoría de la inteligencia de Jeff Hawkins.

Principio de especificidad de las conexiones

En el seno de la neurociencia existe una tensión entre la plasticidad de las conexiones neuronales y el *principio de especificidad de las conexiones* enunciado por Ramón y Cajal. Éste contiene dos consideraciones. La primera es que no hay continuidad citoplásmica entre las neuronas, al contrario de la creencia de su contemporáneo Camillo Golgi, que estaba convencido de que el tejido nervioso era continuo. La segunda es que las neuronas no se comunican indiscriminadamente ni forman redes aleatorias (Kandel, Schwartz & Jessell, 1995, p. 25). Por tanto, por un lado tenemos que las redes de neuronas son indudablemente plásticas, pero por el

otro no es menos cierto que no se unen entre sí de manera aleatoria. Esto implica que para producir una IA subsimbólica mediante la construcción o simulación de redes de neuronas no bastaría con crear miles o millones de neuronas artificiales y conectarlas de manera aleatoria. La experiencia moldea las conexiones como establece el principio de Hebb, pero además la configuración inicial de las redes está determinada por factores genéticos que es indispensable conocer.

En el cerebro humano hay 10^{11} neuronas unidas por 10^{14} conexiones que son producto de la interacción de factores genéticos y ambientales (Ibíd., p. 43). Los mecanismos mediante los cuales participan los factores genéticos todavía no se conocen con precisión. Se sabe que el sistema nervioso de los vertebrados se desarrolla a partir del *ectodermo*, que es la capa más externa del embrión temprano. El tejido ectodérmico que recibe señales químicas del mesodermo se transforma en la *placa neural*, mientras que el resto acabará formando la piel. En el siguiente estadio la placa neural se diferencia en los dos tipos de células que hay en el sistema nervioso: neuronas y células gliales. Las neuronas del sistema nervioso periférico derivan de unas células de la placa neural llamadas *células de la cresta neural*. Por su parte, las neuronas del sistema nervioso central proceden de las *zonas ventriculares del neuroepitelio*. En ambos casos, las neuronas se generan en una zona de origen desde la cual tienen que migrar hasta su destino. Para moverse utilizan unas extensiones llamadas *filopodios* con las que se agarran a la glia circundante.

La cuestión fundamental en este enorme proceso migratorio es saber cuáles son los mecanismos utilizados por las neuronas para desplazarse hasta el destino correcto. La hipótesis más plausible es la de la *afinidad química* (Ibíd., p. 102). Según ésta, las neuronas adquieren en un momento temprano de su desarrollo unos marcadores moleculares distintivos que son afines a ciertas sustancias y repelen otras. Descubrir los detalles del funcionamiento de este mecanismo de afinidad química aplicado a 10^{11} neuronas y 10^{14} conexiones neuronales es una tarea que todavía está por hacer, y que se presenta como una de las más difíciles en la Historia de la ciencia. Al tratarse de una empresa muy grande, es razonable prever que necesitará mucha

financiación. Dado que estamos en la etapa profesional de la ciencia (Woolgar, 1988, p. 30), los proyectos como éste requieren de la financiación privada o bien de una financiación pública excepcionalmente cuantiosa. Sin embargo, la financiación privada se mueve con un único objetivo, que es el rendimiento económico de su inversión, y no parece que la descripción detallada de los factores genéticos de la constitución del sistema nervioso pueda devolver a corto o medio plazo ese tipo de rendimiento. El capital de riesgo orientado a la investigación y el desarrollo ha demostrado en las últimas décadas que prefiere otro tipo de aventuras más solventes, como por ejemplo Apple, una compañía basada desde sus inicios en el robo a sus competidores y en la explotación hasta el suicidio, literalmente, de sus trabajadores.

Anatomía y fisiología de la neurona

Una vez expuesto a grandes rasgos el sistema nervioso, vamos a examinar las neuronas con cierto detalle. Así como para evaluar la viabilidad del proyecto de investigación de la IA simbólica es preciso haber examinado antes las características del paradigma cognitivista en el que se desarrolla, para evaluar la viabilidad del proyecto de investigación de la IA subsimbólica es indispensable haber comprendido previamente el funcionamiento del sistema nervioso humano. Es una tarea que puede parecer alejada de la filosofía, que es el género del presente estudio, pero en realidad los conceptos de la neurofisiología son la base sobre la cual muchos filósofos como Descartes o Hobbes han construido sus teorías más especulativas. Filosofar sobre la mente en nuestros días sin tener en cuenta el acervo de conocimiento proporcionado por la neurociencia es una práctica anacrónica por desgracia muy extendida que se cuenta entre las causas del descrédito de la filosofía.

Una neurona típica tiene cuatro regiones morfológicas: cuerpo celular, dendritas, axón y terminales presinápticas (Ibíd., p. 23). El *cuerpo celular* o *soma* es el centro metabólico de la neurona. De él salen dos tipos de prolongaciones: las dendritas y el axón. Las *dendritas* son estructuras ramificadas que sirven para recibir

señales de otras neuronas. Cada neurona suelen tener varias dendritas, mientras que lo habitual respecto del axón es tener sólo uno. El *axón* es una prolongación que nace en una zona del soma denominada cono de arranque axónico y que se extiende hasta que las *terminales presinápticas*, situadas en su extremo, hacen contacto con otras células. Si la célula objetivo es una neurona, las terminales postsinápticas con las que se contacta axón pueden estar ubicadas no sólo en las dendritas, sino también en el soma o en el axón. En virtud del *principio de polaridad dinámica* establecido por Ramón y Cajal, la corriente eléctrica se propaga a lo largo de las neuronas en un solo sentido, que va desde las dendritas hacia las terminales presinápticas.

Hay dos tipos de corriente eléctrica en una neurona: potenciales electrotónicos y potenciales de acción. Los *potenciales electrotónicos* son corrientes despolarizantes o hiperpolarizantes que se generan en las terminales postsinápticas ubicadas en los extremos de las dendritas y que se desplazan hasta llegar a la *zona de activación* del soma. La zona de activación suele coincidir con el cono de arranque axónico. Si la sumación espacial y temporal de los potenciales electrotónicos despolariza lo suficiente la zona de activación hasta superar una magnitud denominada *umbral de disparo*, entonces se produce una *punta* o *potencial de acción*, que es una señal eléctrica despolarizante que se propaga por el axón hasta llegar a las terminales presinápticas. La unión de las terminales presinápticas de una neurona con las terminales postsinápticas de la siguiente célula se denomina *sinapsis*. En cuanto a su morfología, las sinapsis entre neuronas pueden ser axodendríticas, axosomáticas y axoaxónicas, es decir, de axón a dendrita, de axón a soma y de axón a axón. Por lo general, en el sistema nervioso central las *axodendríticas* son excitatorias, las *axosomáticas* son inhibitorias y en las *axoaxónicas* el axón presináptico modula la sinapsis del axón postsináptico. En cuanto a las sustancias utilizadas para la comunicación intercelular, las sinapsis pueden ser químicas o eléctricas. En el caso de las *eléctricas*, el potencial de acción se propaga directamente a la célula postsináptica perdiendo bastante amplitud, mientras que en las sinapsis *químicas*, que son las más abundantes en todo el sistema nervioso, el potencial de acción despolariza la terminal

presináptica, provocando la liberación de unas sustancias químicas llamadas *neurotransmisores* que son captadas por las terminales postsinápticas de la siguiente neurona. La recepción de neurotransmisores puede dar lugar a la generación de potenciales electrotónicos, a partir de los cuales el ciclo se repite.

La amplitud, la duración y la velocidad son parámetros constantes de los potenciales de acción de una neurona, mientras que la frecuencia y el número de ellos que se dispare son variables que dependen de la duración y de la magnitud de la despolarización de la zona de activación. Cuanto más se haya sobrepasado el umbral de disparo, mayor será la frecuencia de los trenes de potenciales de acción, y a mayor frecuencia y duración de la despolarización, mayor será el número de potenciales de acción disparados. Si el umbral no es sobrepasado, entonces no se produce ningún potencial de acción. Esta respuesta binaria de todo o nada característica de la zona de activación actúa como una puerta lógica que devuelve un 1 o un 0 en función de las entradas recibidas. El médico británico Charles Sherrington calificó esta capacidad celular de decisión como la habilidad más fundamental del cerebro (Ibíd., p. 222). En ella no existen resultados intermedios entre el 1 y el 0 debido a que la amplitud de los potenciales de acción de una neurona es siempre la misma. El umbral se sobrepasará en mayor o menor medida dependiendo de los potenciales electrotónicos recibidos en la zona de activación. Los *potenciales electrotónicos hiperpolarizantes* alejan el potencial de membrana del umbral de disparo, y por tanto tienen un efecto inhibidor del disparo de potenciales de acción, mientras que los *potenciales electrotónicos despolarizantes* acercan el potencial de membrana al umbral de disparo, y por tanto su efecto es excitatorio. En las sinapsis eléctricas los potenciales electrotónicos son siempre despolarizantes, ya que son propagaciones directas de la señal del potencial de acción de la neurona anterior, y los potenciales de acción son siempre despolarizantes. En cambio, en las sinapsis químicas los potenciales electrotónicos producidos pueden ser hiperpolarizantes o despolarizantes. Una sinapsis química sólo puede producir uno de esos dos efectos: hiperpolarización o despolarización, de modo que en el sistema nervioso hay sinapsis *inhibitorias* y *excitatorias* respectivamente.

En las sinapsis químicas la despolarización de los potenciales de acción produce en las terminales presinápticas la fusión de *vesículas sinápticas* con la membrana, de forma que las vesículas liberan los neurotransmisores que llevan en su interior. Los neurotransmisores recorren una pequeña distancia hasta llegar a las terminales postsinápticas de la neurona siguiente, donde son captados por unas estructuras denominadas *receptores*. Al recibir los neurotransmisores, los receptores abren o cierran unos *canales iónicos* que permiten el intercambio de iones entre el interior y el exterior de la neurona. En función del tipo de los iones intercambiados, el efecto es una hiperpolarización en el caso de las sinapsis inhibitorias y una despolarización en las excitatorias. Las dendritas están completamente cubiertas de canales iónicos activables por voltaje que propagan la hiperpolarización o despolarización producida en los receptores sinápticos a lo largo de ellas hasta llegar a la zona de activación. El axón, por su parte, está intermitentemente cubierto de canales iónicos activables por voltaje, pero sólo permiten un intercambio iónico despolarizante, razón por la cual los potenciales de acción son siempre despolarizantes. En las terminales postsinápticas los receptores pueden abrir sus canales iónicos asociados de manera directa o indirecta. Los que lo hacen de manera directa están integrados en la propia estructura del canal iónico, y se denominan *ionotrópicos*, mientras que los que lo hacen de manera indirecta están ligeramente separados de sus canales iónicos correspondientes y pueden ser de dos tipos: *metabotrópicos* y de la *familia de la tirosina quinasa*, siendo los metabotrópicos los mejor conocidos en la actualidad.

Los receptores directos producen acciones sinápticas rápidas que duran sólo unos milisegundos, a diferencia de los indirectos, cuyas acciones sinápticas son lentas pero a cambio pueden durar desde varios segundos a algunos minutos. Para evitar su acumulación indefinida, los neurotransmisores son eliminados de la hendidura sináptica mediante tres mecanismos: *recaptación*, *difusión* y *degradación enzimática*. Sin embargo, hasta que se produce su eliminación automática, permanecen en la hendidura sináptica un tiempo durante el cual suman su efecto al de los neurotransmisores recibidos posteriormente, de modo que actúan como una especie

de memoria química que, en función de lo ocurrido anteriormente, condiciona las comunicaciones interneuronales siguientes. La farmacología aprovecha esta característica de las sinapsis químicas. Por ejemplo, el famoso Prozac, que no es más que una marca comercial de la fluoxetina, es un serotoninérgico que actúa inhibiendo la recaptación de la serotonina, la cual es un neurotransmisor presente en las sinapsis de circuitos neuronales asociados con la sensación de felicidad.

Existen también otros importantes mecanismos de memoria a nivel celular, como por ejemplo la *potenciación postetánica*, que consiste en una acumulación en la terminal presináptica de iones Ca^{2+}, los cuales producen la fusión de las vesículas sinápticas con la membrana, de manera que aun en ausencia de potenciales de acción, la neurona presináptica sigue liberando neurotransmisores durante minutos e incluso durante toda una hora. La potenciación postetánica se produce tras una *estimulación tetánica*, que es aquella con una alta frecuencia de potenciales de acción. Según el principio de Hebb, cuanto más dispara una neurona sobre otra, más fuerte se hace la conexión entre ambas. Esto se debe a mecanismos de memoria como los que acabamos de describir, y también a la liberación de factores de crecimiento por parte de la neurona postsináptica que aumentan la cantidad de neurotransmisores producidos y liberados por la presináptica en el futuro.

4.4. Emergentismo y reduccionismo

En virtud de lo expuesto, resulta evidente que durante el último siglo la neurociencia ha demostrado ser una disciplina madura que ha progresado con la solidez propia de cualquier otra ciencia natural, mientras que la psicología todavía se encuentra en el estadio kuhniano de la pseudociencia debido a que no ha conseguido consolidar el paradigma cognitivista como un marco estable en el que desarrollar una heurística positiva exitosa y acumulativa. Esta disparidad ha llevado a algunos filósofos y científicos a creer que la psicología es prescindible. Tal es el caso del filósofo pragmatista Richard Rorty, quien afirma que: «Los psicólogos deberían ser más

mecanicistas en vez de menos, deberían recorrer el camino que lleva de lo mental a lo neurofisiológico» (Rorty, 1979, p. 217). En opinión de Rorty, el estudio de las representaciones mentales no es necesario para entender la conducta. Está convencido de que la psicología ha sido un producto azaroso de las contingencias históricas, pues «si la fisiología fuera más simple y más obvia de lo que es, nadie sentiría la necesidad de la psicología» (Ibíd., p. 237). Rorty propone que la epistemología debería contemplar el examen de sólo dos elementos. Por un lado, la biología del cuerpo humano, y por el otro, las condiciones históricas y culturales que afectan a ese cuerpo. Este tipo de enfoque se denomina *reduccionismo*, dado que propone la reducción de la mente al cerebro, es decir, del plano representacional en el que es necesario operar con la noción de causalidad final al plano biológico en el que todo ha de ser explicable en términos de causalidad eficiente.

En esta última sección del presente capítulo vamos a exponer dos argumentos en contra del reduccionismo. El primero se basa en una analogía de John Searle, mientras que el segundo es una apreciación de Hillary Putnam sobre la intransitividad de las explicaciones que impide reducir una explicación a otras de nivel inferior sin que en el proceso se pierda algún conocimiento esencial. Ambos nos servirán para defender una teoría emergentista de la mente y para señalar la necesidad de que la neurociencia trabaje junto con la psicología sin sustituirla.

Emergentismo

Según Searle: «Todas nuestras experiencias conscientes se explican por el comportamiento de las neuronas y son propiedades emergentes del sistema de neuronas» (Searle, 1996, p. 658). Y a continuación explica lo que son las propiedades emergentes: «Una propiedad emergente de un sistema es una propiedad que se explica por el comportamiento de los elementos de dicho sistema pero no pertenece propiamente a ninguno de sus elementos ni puede explicarse simplemente como la suma de las propiedades de dichos elementos. La liquidez del agua es un buen

ejemplo: efectivamente, el comportamiento de las moléculas de H_2O explica la liquidez del agua, sin embargo ninguna de las moléculas individuales es líquida» (Ibíd., p. 659).

Recordemos las nociones de explicación morfológica y explicación sistemática que vimos al comienzo de este capítulo. Las explicaciones morfológicas son aquellas que explican un fenómeno apelando a la estructura y a las propiedades de las partes que lo componen (Haugeland, 1978, p. 246). Las explicaciones sistemáticas son similares a las morfológicas, con la diferencia de que el fenómeno explicado surge de la interacción de las partes (Ibíd., p. 247). Cuando Searle dice que la liquidez no puede explicarse como la suma de propiedades de los elementos del agua quiere decir que no es explicable morfológicamente, pues la liquidez es una propiedad que emerge del sistema formado por la interacción entre las moléculas de H_2O. En una molécula de H_2O el átomo de oxígeno tiende a atraer electrones y por tanto lleva una pequeña carga negativa, mientras que los átomos de hidrógeno tienden a perder electrones y en consecuencia llevan una pequeña carga positiva (Kandel, Schwartz & Jessell, 1995, p. 116). Gracias a estas propiedades individuales, cuando se juntan varias moléculas de H_2O emergen entre ellas los denominados *puentes de hidrógeno*, que son los que producen la propiedad de la liquidez. La liquidez, por tanto, es una propiedad emergente de las moléculas de agua, como dice Searle, a lo cual nosotros añadimos que también emerge por la interacción con las condiciones ambientales, pues la liquidez depende tanto de las propiedades iónicas de los átomos de las moléculas de agua como de unas condiciones ambientales adecuadas de presión y temperatura.

En el caso de la mente sucede algo similar. Marvin Minsky, uno de los padres fundadores de la IA, escribió una sentencia proverbial al respecto: «Las mentes son sólo lo que hacen los cerebros» (Minsky, 1985, p. 287). Esta concepción sobre la relación entre la mente y el cerebro se denomina *emergentismo*, y es la que nosotros defendemos, respaldados por nombres tan ilustres como Charles Darwin, Santiago Ramón y Cajal, Donald Hebb, Vernon Mountcastle, Mario Bunge, John Searle, José Ferrater Mora y Antonio Damasio (García, 2001a, p. 278), por citar sólo a autores que ya hemos mencionado anteriormente. Renunciar por principio al estudio de la mente,

que es lo que propone Rorty, sería como renunciar a estudiar las propiedades que surgen de la interacción entre las moléculas de agua, entre ellas la liquidez. Estaríamos entonces renunciando a examinar un fenómeno que, además de ser interesante por sí mismo, contribuye a realizar nuevos descubrimientos sobre la materia y las condiciones ambientales de las que él emerge. Del mismo modo, los fenómenos mentales estudiados por la psicología tienen un interés intrínseco, pero también ayudan a comprender el funcionamiento del cerebro y el efecto de las condiciones ambientales sobre la cognición.

La propuesta reduccionista de Rorty es justo la fotografía en negativo de la pretensión cognitivista de reclamar un nivel de análisis de la mente aislado en buena medida del conocimiento del cerebro y de los factores ambientales más complejos. Ambas son planteamientos eliminativistas tan defectuosos como el conductismo, en tanto que comparten por principio la mutilación del fenómeno real de la cognición. Utilizando la analogía del agua, el paradigma cognitivista estaría reclamando su derecho a estudiar la liquidez ignorando lo que la física y la química dicen sobre los átomos y sobre la temperatura y la presión atmosférica. Lo más que se puede obtener desde ese enfoque es un conocimiento práctico sobre la liquidez, que es el que tienen los psicólogos, los cuales en su mayoría salen de las universidades con conocimientos que muchas veces funcionan, pero no saben por qué. Los psicólogos suelen ser al estudio de la mente lo que los cocineros son a la química. En el ejercicio de la profesión raros son los casos como el de Howard Gardner, un psicólogo preocupado por todo lo que concierne a la mente, desde la filosofía hasta la neurociencia, pasando por la lingüística y la antropología. A diferencia de él, la mayoría de sus colegas se desentiende del estudio de las causas profundas y se limita al conocimiento de la técnica. Después, la sociedad les encomienda tareas de suma importancia tales como diseñar planes educativos y evaluar el grado de peligrosidad de la población penitenciaria para decidir su excarcelación. Las consecuencias son de sobra conocidas, pero no son culpa de ellos, sino de quienes les confían esas tareas, pues es como pedirle a un cocinero que prepare un barril de nitroglicerina.

Reduccionismo

Respecto al reduccionismo, Hillary Putnam lo rechaza señalando que es falso que las ciencias de alto nivel como la psicología y la sociología puedan ser reducidas a leyes de ciencias de bajo nivel, como la biología, la química y, en última instancia, la física de partículas. El filósofo norteamericano argumenta que «del hecho de que el comportamiento de un sistema pueda ser *deducido* de su descripción como un sistema de partículas elementales no se sigue que pueda ser *explicado* a partir de esa descripción» (Putnam, 1973a, p. 205). Es una afirmación que puede parecer contraria a la intuición porque se opone a dos creencias muy extendidas. La primera es que la deducción de un fenómeno equivale a explicarlo, y la segunda, que las explicaciones son transitivas. Sin embargo, ambas creencias son falsas, dice Putnam. En cuanto a la primera, supongamos que un hecho F se deduce de G e I, siendo G una explicación genuina e I algo irrelevante. Todo el mundo consideraría que sólo G es explicativa de F. Pero supongamos que G e I son fusionadas en una sola proposición H matemáticamente equivalente a ellas de modo que a partir de G no se pueden separar H e I. En este caso H no sería una explicación de F, pero F sería deducible de H. Así ilustra Putnam la diferencia entre deducir y explicar.

Acerca de la segunda creencia, Putnam indica que en realidad las explicaciones no son transitivas. Esto se debe a que la explicación de una explicación contiene información que es irrelevante para lo que queremos explicar, y además contiene la información relevante en una forma que puede ser imposible de distinguir de la irrelevante, como sucede en la explicación G. Un ejemplo lo encontramos en la relación entre la neurociencia y el capitalismo. La neurociencia ofrece explicaciones válidas sobre el funcionamiento del cerebro, pero dados los conocimientos de la neurociencia sobre la estructura del cerebro y del sistema nervioso, uno no puede deducir que el sistema de producción capitalista existirá, pues «las mismas criaturas pueden existir en un sistema de producción precapitalista, o en uno feudal, o en uno

socialista, o en otros de muchos tipos. Las leyes de la sociedad capitalista no pueden ser deducidas de la suma de las leyes de la física más una descripción del cerebro humano: dependen de "condiciones fronterizas" que son *accidentales* desde el punto de vista de la física pero esenciales para describir una situación como un sistema capitalista. En resumen, las leyes del capitalismo tienen una cierta *autonomía* frente a las leyes de la física: tienen una *base* física (los hombres tienen que comer), pero no pueden ser deducidas de las leyes de la física. Son compatibles con las leyes de la física, pero también lo son las leyes del socialismo y las del feudalismo» (Ibíd., p. 209).

Por si este ejemplo no resultara suficientemente claro, Putnam pone a la evolución como otro caso de intransitividad de las explicaciones. Señala que la evolución depende de la microestructura contenida en el genotipo de los organismos, pero también depende de condiciones ambientales como la presencia de oxígeno que son accidentales desde el punto de vista de la física y la química: «Las leyes de las disciplinas de más alto nivel son deducibles a partir de las leyes de las disciplinas de más bajo nivel junto con "hipótesis auxiliares" que son accidentales desde el punto de vista de las disciplinas de más bajo nivel. Y la mayoría de la estructura en el nivel de la física es irrelevante desde el punto de vista de las disciplinas de más alto nivel; sólo algunas características de esa estructura, tales como el genotipo en el caso de la evolución, son relevantes, y son especificadas por las disciplinas de más alto nivel, no por las de menor nivel» (Ibíd., p. 209).

El argumento de Putnam contra la transitividad de las explicaciones se basa en la definición de modelo y de teoría que vimos antes. Un modelo o una teoría es siempre una simplificación de una parcela de la realidad, y opera seleccionando o sobrevalorando algunos rasgos (García, 2001a, p. 77). En las teorías que explican el capitalismo tienen cabida algunos fenómenos biológicos, pero no todos. Hay fenómenos que son ignorados por ser insignificante su importancia para la teoría. En el caso del tiro parabólico se desprecia el color del proyectil. Es cierto que el color influye, pero lo hace en una medida tan pequeña que es dejado fuera de la teoría. En cambio, el color sí resultaría un factor relevante en una teoría no sobre el movimiento,

sino sobre la observación de los cuerpos en la oscuridad. En ese caso, cuanto más cercano al blanco fuese el color del proyectil, más visible sería. Recordemos que una de las deficiencias lógicas de la inducción como estrategia de descubrimiento es precisamente que, si procediéramos por pura inducción, todos los factores que rodean al fenómeno observado deberían ser incluidos, desde el color de las paredes del laboratorio hasta el olor de la colonia del científico, de manera que la inducción resultante sería inmanejable. Las teorías se elaboran en su mayoría abductivamente, seleccionando ciertos rasgos de entre una totalidad fenoménica potencialmente infinita. Y, como dice Einstein, no hay ningún método que indique cuáles de esos rasgos deben ser seleccionados ni cómo deben ser relacionados entre sí (Einstein, 1955, p. 131). En el contexto de descubrimiento el investigador realiza su elección motivado por causas teóricas y metateóricas. Y después, en el contexto de justificación operan también causas de ambos tipos. Las metateóricas proceden en parte de todo el trasfondo cultural que opera consciente e inconscientemente sobre las mentes de los miembros de la comunidad científica.

Un positivista extremo, movido por la admiración hacia el estilo explicativo nomológico-deductivo de la física y por el afán reduccionista de su doctrina, podría llegar a defender que todo en el universo, desde la evolución de las especies hasta el capitalismo, es reducible a una explicación en términos de la física de partículas. Si creemos en el determinismo, quizás tuviera razón. Dejando a un lado las paradojas lógicas a las que daría lugar, una teoría de ese tipo serviría para deducir con exactitud cualquier instante del pasado y del futuro, tal y como lo haría el demonio de Laplace. Ahora bien, todas sus explicaciones estarían formuladas en una teoría del nivel más bajo. La predicción de dónde caería un proyectil que describiese una parábola sería la más exacta imaginable, a un nivel de detalle cuántico que nos indicaría la posición de cada una de las partículas subatómicas del proyectil. Pero eso no sería económico.

Las teorías científicas han de ser *económicas* no sólo a nivel ontológico como exige el principio de parsimonia, sino también a nivel epistemológico, es decir, en el nivel de la *comprensión* de los fenómenos que describen. Para deducir el lugar de caída

de las partículas del proyectil haría falta una cantidad demencial de ciclos de computación, y aplicar después una simplificación para que la computadora devolviera los datos en un formato de más alto nivel asequible para la mente humana, como el de las coordenadas de latitud y longitud. Si sólo devolviera los datos de más bajo nivel, el ser humano interesado en conocer el lugar de la caída del proyectil tendría que dedicar varios años, o quizás toda su vida, a leer una pantalla que no cesa de imprimir números. Por tanto, las explicaciones de alto nivel seguirían siendo necesarias. La predicción de la física de partículas no sería más que una duplicación exacta de la realidad, a la cual habría que aplicarle teorías de alto nivel para extraer los conocimientos que contiene. En la película de ciencia ficción *The Matrix* (1999), mencionada en el primer capítulo, los personajes podían saber lo que ocurría en un falso mundo creado por ordenador leyendo su descripción en tiempo real en una pantalla llena de símbolos cambiantes. Es de suponer que esos símbolos eran abstracciones de alto nivel y solamente sobre algunos hechos en particular, pues si hubieran sido descripciones de los movimientos de todas las partículas del universo virtual creado por Matrix, entonces habrían tardado una eternidad en entender algo tan simple como que había un gato sobre una alfombra.

En conclusión, al menos desde el enfoque emergentista que defendemos, el reduccionismo de la mente al cerebro es un error. Las explicaciones no son transitivas debido a que los modelos y las teorías son siempre abducciones simplificadoras de la realidad que valoran o desprecian a un cierto nivel fenómenos que en cambio a otros niveles sí son relevantes. Deducir y explicar no son la misma cosa. Entender un fenómeno, ya sea natural o social, consiste siempre en distinguir lo relevante de lo irrelevante y en descubrir las relaciones entre los factores relevantes que lo producen tal y como se presenta desde un determinado punto de vista. En función del fenómeno y del punto de vista, esto a veces puede hacerse en un estilo explicativo nomológico-deductivo, pero otras veces no se puede, y hay que optar por un estilo explicativo morfológico o sistemático. La mente es un fenómeno muy complejo que emerge de la interacción de factores biológicos expresados habitualmente en el estilo nomológico-

deductivo y de factores sociales que, en el mejor de los casos, sólo parecen ser expresables en un estilo sistemático. La obsesión del cognitivismo por la formalización de los procesos mentales en ecuaciones lógico-matemáticas para hacerlos manipulables por la computadora electrónica ha dado lugar a lo que Howard Gardner denomina la *paradoja computacional*.

La paradoja computacional consiste en que la metáfora computacional, más que haber servido para entender cómo funciona la cognición, ha servido para descubrir cómo de hecho *no* funciona la cognición. Los cognitivistas han cometido un exceso metodológico similar al de los conductistas: reformular el objeto de estudio para adecuarlo a las herramientas de las que se dispone (Gardner, 1985, p. 414). Dado que los procesos cognitivos humanos más complejos no son asequibles para el modelo computacional, los cognitivistas han optado por defender la metáfora computacional, tal y como les exige la heurística negativa, y despreciar ese tipo de procesos. Los científicos cognitivos han cometido el error de «excluir del examen factores nada triviales, como el papel del contexto circundante, los aspectos afectivos de la experiencia y la repercusión de los factores culturales e históricos en el comportamiento y la conducta. [...] Creo que en última instancia la ciencia cognitiva tendrá que abordar estos factores» (Ibíd., p. 415). Gardner continúa diciendo: «Paradójicamente, gran parte de los mejores trabajos de la ciencia cognitiva se llevaron a cabo como si sólo existiese el nivel de la representación mental. Por ejemplo, en el caso del lenguaje (más específicamente, de la gramática), la brillante obra de Chomsky y sus seguidores no hace referencia alguna a las condiciones reales del cerebro y de la cultura circundante, y podría sustentarse con independencia de dichas condiciones» (Ibíd., p. 419).

Ciertamente, elaborar una psicología que contemple todos los factores que producen ese fenómeno llamado mente, desde la biología del cuerpo humano hasta la cultura, se presenta como una tarea harto difícil. Pero ahí es donde hay que buscarla. Adoptar un enfoque eliminativista como hace el cognitivismo es hacer lo mismo que el borracho del chiste, que en plena noche cerrada estaba buscando algo a la luz de una

farola. Entonces se le acerca un hombre y le pregunta: "¿Está seguro de que se le ha perdido por aquí?". Y el borracho contesta: "No. Estoy seguro de que se me ha perdido por allí, pero es que allí está oscuro, y aquí hay luz".

Resumen

El paradigma cognitivista entraña al menos dos graves errores contenidos en sus dos supuestos nucleares. En primer lugar, la tesis internalista proclama que es legítimo proceder metodológicamente como si el dualismo de sustancias cartesiano fuera verdadero y por tanto la mente fuese analizable separada del cerebro en particular y del cuerpo en general, de manera análoga a la creencia errónea de que el software es analizable con independencia del hardware. Y segundo, despreciar aquellos fenómenos de la vida psíquica, como las emociones y los factores ambientales complejos, que por no ser expresables en proposiciones de un lenguaje formal manipulable por una computadora tal y como exige la metáfora computacional, alejan a la psicología de la pretensión positivista de convertirla en una ciencia que, a pesar de su condición híbrida de disciplina biosocial, posea un rendimiento instrumental tan elevado como el de las ciencias puramente naturales, cuyos modelos de las parcelas de la realidad que describen son computables. Cuando analicemos la Historia de la IA en el capítulo sexto veremos que la IA se adscribió en sus orígenes al paradigma cognitivista, arrastrando en consecuencia estos dos errores que durante décadas han impedido tanto su avance como el de la psicología.

Por el momento, en el próximo capítulo vamos a examinar algunas teorías contemporáneas de la inteligencia, pues para evaluar las condiciones de posibilidad técnicas de la IA fuerte, que es el objetivo del capítulo séptimo, necesitamos saber primero qué es la inteligencia, profundizando más allá de la definición provisional de Donald Michie que hemos manejado hasta ahora.

5. Teorías contemporáneas de la inteligencia

La inteligencia es un concepto extremadamente escurridizo. En 1986 Robert Sternberg y Douglas Detterman editaron *What is intelligence?*, una recopilación de veinticuatro artículos firmados por algunos de los psicólogos más prestigiosos del momento con la que pretendían esclarecer el estado de la cuestión sobre la inteligencia en esa fecha y compararlo con las conclusiones de un simposio clásico de 1921 que perseguía el mismo objetivo. Paul Baltes afirmaba en su correspondiente artículo que «el núcleo teórico de la inteligencia [...] no puede ser delineado de una manera clara y ampliamente aceptable» (Baltes, 1986, p. 23). Sandra Scarr opinaba que «hay tantas respuestas a la pregunta "¿qué es la inteligencia?" como propuestas para formularla» (Scarr, 1986, p. 117). Y el propio Sternberg reconocía en el capítulo introductorio que «los psicólogos no pueden ni siquiera ponerse de acuerdo sobre qué es la inteligencia» (Sternberg, 1986, p. vii).

En la sección final dedicada a las conclusiones Detterman resolvía que, a la luz de la divergencia entre las opiniones de sus colegas, era evidente que el estado de la cuestión había cambiado poco en los 65 años transcurridos desde el simposio de 1921. La misma disparidad de criterios sobre la definición de la inteligencia se mantenía en 1986, y aún perdura en la actualidad (Davidson & Kemp, 2011, p. 58). Según Detterman, esta pluralidad de puntos de vista no debería entenderse como una señal de que nadie sabe en realidad de qué está hablando, sino más bien como un signo de vigor en una empresa científica que todavía se encuentra en su periodo formativo (Detterman, 1986, p. 164). Lo cierto es que la falta de consenso sobre asuntos fundamentales como éste es constante en la psicología, y es la causa de que esta disciplina no sea una ciencia madura organizada en torno a un solo paradigma.

Dado que no existe una definición universalmente aceptada de la inteligencia, vamos a dedicar el comienzo de este capítulo a esbozar y a argumentar en favor de la teoría contemporánea de la inteligencia que a nuestro juicio es la más acertada: la *teoría de las inteligencias múltiples*, o *teoría IM*, de Howard Gardner. Más adelante volveremos a ella para examinarla en profundidad. Conviene recordar aquí lo que dijimos en el primer capítulo: nuestro interés se centra en la reproducción artificial de la inteligencia humana, no de la inteligencia en general. Aunque los animales también tienen cierto grado de inteligencia y por tanto pueden aplicárseles muchas de las observaciones que vienen a continuación, cuando hablemos de la inteligencia sin más nos estaremos refiriendo siempre a la de los seres humanos.

La teoría de las inteligencias múltiples

Jacqueline Goodnow, psicóloga de la Universidad Macquarie de Sídney, afirma que: «En lugar de pensar en la inteligencia como una cualidad residente en el individuo [...] propongo que la consideremos como un juicio o atributo, comparable con los juicios cotidianos que hacemos sobre la gente acerca de su atractivo físico, ingenio, amistosidad o timidez. En todos estos juicios somos conscientes de la relatividad de los cánones, de la fiabilidad del consenso sobre los casos extremos y de la posibilidad del sesgo. ¿Por qué deberían ser diferentes los juicios sobre la inteligencia?» (Goodnow, 1986, p. 85). A nuestro parecer, Goodnow acierta al advertir que la inteligencia es, de hecho, un atributo que en la vida cotidiana predicamos de las personas de acuerdo a convenciones sociales.

Observemos que el canon de belleza depende de factores culturales. En Occidente no hace mucho que se consideraba que una persona bella debía tener la piel muy blanca, pues era señal de que no tenía necesidad de trabajar y por tanto no le daba el sol. Hoy en día el trabajo suele realizarse en lugares cerrados, como las oficinas, y por tanto el canon se ha invertido, de forma que entre las personas de raza caucásica se considera bello el tener la piel morena, pues es indicador de que se

dispone de tiempo libre para solazarse. De manera análoga, Noam Chomsky, por ejemplo, es considerado un hombre inteligente en las sociedades occidentales, contextos donde se valora mucho el conocimiento académico que él posee, pero en cambio es probable que en una tribu de la Polinesia fuera considerado deficiente mental, pues es de suponer que un hombre como él no entrena a menudo las destrezas intelectuales que en una sociedad así son valoradas, tales como la habilidad para distinguir con rapidez a un animal peligroso oculto en la maleza.

Sin cambiar de cultura, dentro de Occidente también observamos que se han producido cambios en la valoración de las inteligencias a través del tiempo. En el capítulo segundo comentamos que en la Baja Edad Media las matemáticas tenían una reputación dudosa, ya que eran consideradas un arte oscura propia de clérigos y astrólogos (Guijarro & González, 2010, p. 72), mientras que en la actualidad la destreza en el cálculo aritmético y geométrico es una de las habilidades más apreciadas socialmente en todos los ámbitos, desde la actividad científica hasta el mundo laboral ordinario. Dentro de esta concepción contextualizada de la inteligencia, uno de los psicólogos contemporáneos más ilustres es Howard Gardner, a quien ya nos hemos referido largamente en el capítulo anterior.

Como investigador crítico que se rebela contra el eliminativismo y el aislacionismo característicos de los supuestos nucleares del paradigma cognitivista, Gardner propone un concepto integral de inteligencia que contempla desde el nivel biológico hasta el cultural. Según él, la *inteligencia* es «la capacidad para resolver problemas, o para elaborar productos que son de gran valor para un determinado contexto comunitario o cultural» (Gardner, 1993, p. 27). Frente a los teóricos que, como Robert Sternberg, consideran que sólo existe una inteligencia *unitaria* que se aplicaría indistintamente a todo tipo de situaciones (Sternberg, 1999, p. 410), Gardner sostiene una concepción *plural* y *modular* de la mente, según la cual hay tantas inteligencias como capacidades básicas para resolver problemas o para elaborar productos de valor cultural. El número de esas inteligencias distinguidas por Gardner ha variado a lo largo de su obra en un número que oscila aproximadamente entre siete

y nueve, pero suele considerarse que las principales han sido siempre estas siete: musical, cinético-corporal, lógico-matemática, lingüística, espacial, interpersonal e intrapersonal. Las dos adicionales, de las que el propio autor no está completamente seguro, son la naturalista y la existencial.

Así como en la belleza reconocemos varios elementos independientes, tales como el bronceado, la sonrisa, la estatura o el peso, en el ámbito intelectual también hay habilidades independientes entre sí, pues observamos que hay individuos que poseen algunas en un grado excepcionalmente alto mientras que en otras son deficientes. La Historia está llena de grandes genios que, por ejemplo, tenían una inteligencia lógico-matemática descollante y que sin embargo eran incapaces de relacionarse socialmente por culpa de una inteligencia interpersonal deficitaria. La inteligencia, por tanto, es un atributo que necesita modificadores, y esto es algo que la teoría de las inteligencias múltiples de Gardner contempla. La definición de la inteligencia de Donald Michie que habíamos manejado hasta ahora, según la cual el intelecto humano destaca por su capacidad de hacer un intento pasable en casi cualquier cosa (Michie, 1974, p. 51), encaja perfectamente con la de Gardner, dado que las siete inteligencias enumeradas por éste agotan exhaustivamente los siete grandes dominios en los que operan todas las competencias intelectuales en el contexto de cualquier cultura.

Ciertamente, la teoría IM de Gardner, que nosotros suscribimos, es discutible, pues de lo contrario no habría tantas opiniones divergentes sobre el tema. Por un lado está el debate entre la concepción unitaria o plural, y por otro, dentro de la concepción plural no hay unanimidad sobre la división de los dominios y lo que éstos deben abarcar. Robert Glaser, por ejemplo, excluye la cognición emocional del ámbito de la inteligencia (Glaser, 1986, p. 79). En tal caso las inteligencias intrapersonal e interpersonal desaparecerían del cuadro, dado que ambas conforman lo que Daniel Goleman denomina *inteligencia emocional*. Sin embargo, desde un enfoque fielmente descriptivo consideramos que es necesario incluir la inteligencia emocional. Los líderes de masas son una clara demostración de que el dominio de las emociones forma parte

de la inteligencia, pues les permite resolver problemas, y en el caso de los artistas es una habilidad imprescindible para la creación de obras de valor cultural. Por otra parte, el caso de Phineas Gage y de los pacientes con lesiones cerebrales similares a la suya constituyen pruebas notorias de que las emociones son necesarias para dirigir la vida de manera inteligente. Respecto de las inteligencias artificiales en sentido fuerte, la gente, conocedora gracias a la ciencia ficción del momento de la imperfectibilidad insalvable de la duplicación del hombre por el hombre, suele creer que las máquinas inteligentes nunca tendrán emociones, pero lo cierto es que la experiencia indica que, muy al contrario, sin emociones no parece posible que una computadora electrónica pueda ser tan inteligente como un ser humano (Franklin, 1995, p. 229).

Más adelante volveremos a la teoría IM de Howard Gardner para examinarla con más detalle. Además, veremos otras dos teorías contemporáneas de la inteligencia: las de Jeff Hawkins y Roger Schank. La de Hawkins nos servirá para entender qué es la inteligencia desde el punto de vista de la neurociencia, que es la disciplina en la que se basa la IA subsimbólica, mientras que la de Schank nos servirá para entender qué es la inteligencia desde el punto de vista de la psicología cognitiva, que es la disciplina en la que se basa la IA simbólica. Antes de eso, vamos a exponer una selección de cinco metáforas de la inteligencia formuladas por un ingeniero informático y cuatro psicólogos. Dado que la inteligencia es un concepto escurridizo, no son pocos los investigadores que recurren al recurso literario de la metáfora para sugerir qué es aquello que no alcanzan a definir con el discurso científico.

5.1. Metáforas de la inteligencia

Son cinco las metáforas que hemos escogido. La primera es de Jeff Hawkins, un ingeniero informático que, desde la teoría de la evolución, afirma que la función definitoria de la inteligencia es que actúa como una *máquina del tiempo* para predecir el futuro. La segunda es de Robert Sternberg, quien compara la inteligencia con el *gobierno de una sociedad*, en lo que es una clara revisión de la tesis platónica de la

semejanza entre las estructuras del alma y de la *polis* (ciudad) ideal. En la tercera el psicólogo John Horn, uno de los creadores de la influyente teoría de la inteligencia CHC, acrónimo de Cattell, Horn y Carroll, compara la inteligencia con un *pudín* (o *pudding*), una analogía que recuerda al modelo atómico del pudín de pasas de John Thomson. La cuarta es de Douglas Detterman, que traza un paralelismo entre las *formas de evaluar la calidad de una universidad* y los enfoques multifactoriales y de factor general dentro de la psicometría. Y finalmente, en la quinta metáfora, Robert Glaser enumera varias características de la inteligencia mostrando su semejanza con una *habilidad atlética* y señalando una distinción muy importante entre competencias naturales y artefactuales a la que nos referiremos a menudo. Si combinásemos todas estas analogías, el resultado sería que la inteligencia es como una máquina del tiempo con la estructura de un pudín y las características de una habilidad atlética que sirve para gobernar el organismo y puede ser evaluada como una institución educativa. Puede parecer una idea confusa, pero en realidad sus partes encajan, dado que cada una de ellas se ocupa de un aspecto de la inteligencia, y no de ésta en su totalidad.

Máquina del tiempo

Desde un enfoque marcadamente corticalista, Jeff Hawkins sostiene que la inteligencia reside en la corteza cerebral. Según él, todo lo que pensamos ocurre ahí, ya sea la percepción, el lenguaje, la imaginación, la matemática, el arte, la música o la planificación (Hawkins & Blakeslee, 2004, p. 55). Hawkins reconoce que el ser humano es mucho más que un organismo inteligente, pero al mismo tiempo señala que su propósito como ingeniero informático y como empresario del sector de las nuevas tecnologías no es producir seres humanos artificiales, sino sólo inteligencias artificiales en el sentido de dobles redoblantes imperfectos. Apelando a los estudios con base experimental de Antonio Damasio ya hemos argumentado que la inteligencia humana depende en buena medida de que la corteza esté conectada con estructuras subcorticales como la amígdala, por ejemplo, que es el centro del sistema límbico,

encargado del procesamiento de las emociones. No obstante, se le puede conceder a Hawkins su punto de vista corticalista, ya que no afecta a su tesis de la inteligencia como un sistema de memoria. La única corrección que habría que hacerle consistiría en fundamentar el funcionamiento de la corteza sobre el resto del sistema nervioso.

Hawkins basa su teoría en la *hipótesis de Mountcastle*, la cual postula que la corteza cerebral utiliza la misma herramienta computacional para todo lo que hace. Es importante advertir que en la hipótesis de Mountcastle no debe entenderse la computación en el sentido convencional, según el cual el procesador de información y el almacén de memoria son estructuras diferentes, sino la computación en el sentido del procesamiento de la información realizado por un *sistema de memoria*. La diferencia estriba en que en el sistema de memoria el procesamiento y el almacenamiento suceden en la misma unidad funcional, que en este caso es la neurona. Toda la corteza procesa información y toda la corteza almacena memoria, dice Hawkins, lo que es una diferencia arquitectónica fundamental respecto de las computadoras electrónicas, que están estructuradas siguiendo el esquema de von Neumann, basado a su vez en el de la máquina analítica de Babbage, que consistía en separar el procesador y el almacén de memoria (Guijarro & González, 2010, p. 297). La herramienta computacional universal que opera en la corteza, dice Hawkins, se dedica a dos tareas, que en realidad son la misma pero en sentidos opuestos: la creación y la aplicación de representaciones invariables. La corteza crea representaciones invariables a partir de la información entrante en el sistema nervioso, y posteriormente las aplica sobre la información entrante en el futuro para generar predicciones. Examinaremos este proceso detenidamente más adelante en este capítulo, pero por ahora sólo nos interesa describirlo a grandes rasgos para entender la metáfora de la máquina del tiempo.

Las representaciones invariables almacenan generalizaciones sobre cómo han sucedido las cosas en el pasado. Ante una situación similar en el futuro, la corteza dispara automáticamente las representaciones invariables asociadas a dicha situación, produciendo un pronóstico de lo que va a suceder. De esta manera, la corteza

cerebral, sede de la inteligencia según Hawkins, funciona como una máquina del tiempo, permitiendo a los organismos adelantarse a los sucesos futuros, y por tanto aumentando sus posibilidades de supervivencia. «La predicción no es sólo una de las cosas que hace nuestro cerebro. Es la función primordial de la corteza cerebral y la base de la inteligencia» (Hawkins & Blakeslee, 2004, p. 109). La selección natural se ha inclinado en favor de los mamíferos debido a que tenemos una corteza cerebral mucho mayor que la de otras clases de vertebrados, como por ejemplo los reptiles (Bustamante, 2007, p. 30). Ésta es una de las razones por las cuales los experimentos en psicología se hacen habitualmente con ratas, y no con lagartijas o cocodrilos. Sin memoria no hay predicción, y sin predicción no hay inteligencia, porque la inteligencia es ante todo predicción. La inteligencia no surgió en el curso de la evolución para que algún día los filósofos disertaran sobre el ser y lo ente, sino para ver el futuro y adelantarse a las situaciones de peligro o beneficio. En su relato de ciencia ficción *El hombre dorado*, el gran escritor Philip K. Dick imagina un animal antropomorfo con la capacidad de ver el futuro de los próximos minutos con total precisión (Dick, 2007, pp. 46 a 76). Un grupo de científicos y militares intenta retener a ese raro espécimen en un recinto de máxima seguridad empleando para ello toda la tecnología a su alcance, pero fracasan en el intento porque es imposible someter a algo o a alguien que sabe lo que va a suceder sin margen de error.

La *prognosis*, o conocimiento del futuro, es el poder supremo, un hecho contemplado por la metáfora de Hawkins y que revela las causas de dos fenómenos tratados en el capítulo anterior. El primero es que los seres humanos aplicamos la inducción como estrategia de anticipación aun a pesar de que es una estrategia no válida desde el punto de vista de la lógica. Lo hacemos porque nuestra corteza cerebral es un sistema de memoria que ha sido modelado por la evolución para dedicarse permanentemente a la anticipación del futuro. El segundo es la obsesión positivista por las explicaciones de estilo nomológico-deductivo. Precisamente se trata de las explicaciones que ofrecen la prognosis más exacta, y por eso son el objetivo prioritario de los científicos y de los fondos de capital que financian los proyectos de

investigación. La capacidad de predicción permite adelantarse al futuro y por tanto prepararse para afrontarlo, o como decía Augusto Comte: prever para proveer. Atendiendo al concepto de inteligencia de Jeff Hawkins, las ciencias de estilo explicativo nomológico-deductivo serían la forma más elevada de inteligencia.

Gobierno de una sociedad

La siguiente metáfora de la inteligencia que hemos elegido es de Robert Sternberg, quien considera que la inteligencia es al individuo lo que el gobierno a una colectividad de individuos (Sternberg, 1986, p. 141). Aunque Sternberg no menciona en ningún momento a Platón, resulta evidente el paralelismo de esta analogía con la trazada por el filósofo griego entre el alma (ψυχή) y la ciudad (πόλις). Decía Platón que el alma encarnada es tripartita, siendo sus partes la razón, el apetito irascible y el apetito concupiscible. La sociedad perfecta debería estar organizada según él a semejanza del alma, dividiéndose respectivamente en una clase gobernante, una de guerreros y otra de productores de bienes materiales (Platón, *La República*, 434-436). Para que una sociedad funcione de la mejor manera, en los gobernantes la parte predominante de sus almas debe ser la razón, en los guerreros, el apetito irascible, y en los productores, el apetito concupiscible.

Por su parte, Sternberg afirma que, así como los gobiernos se dividen en los poderes ejecutivo, legislativo y judicial, la misma división se da en la inteligencia, la cual legisla produciendo reglas sobre cómo es el mundo, ejecuta poniendo en marcha los mecanismos adecuados para afrontar cada situación, y juzga si los resultados obtenidos son aceptables. Cada una de estas tareas se realiza en el seno de una jerarquía, pues en la inteligencia existe una estructura piramidal semejante a la de cualquier gobierno. Cada estrato de la jerarquía ejerce su mandato sobre una parcela, siendo estas parcelas divisibles de muchas formas. Así como un país puede ser dividido en función de diversos criterios como la población, la orografía o el clima, las partes que se pueden distinguir en la inteligencia dependen del objetivo analítico para el que

se realiza la partición. De aquí que no haya consenso en la comunidad investigadora acerca de cuáles son las estructuras que componen la inteligencia. Dichas partes son gobernadas jerárquicamente, como decimos, pero una misma jerarquía puede ser gobernada de diferentes maneras.

Sternberg apunta que el pionero de la psicometría Charles Spearman se dio cuenta a principios del siglo XX de que los diferentes modelos de la inteligencia podían ser considerados análogos a las diferentes formas de gobierno. El esquema clásico de Aristóteles contempla tres: monarquía, oligarquía y democracia, a los cuales cabe añadir la anarquía. No todos funcionan igual de bien. En su *teoría triárquica de la inteligencia* Sternberg propone que las mentes equilibradas funcionan como una oligarquía federada. En cambio, las mentes más desorganizadas se podría decir, en su opinión, que operan con un gobierno anárquico. A su vez, dentro de cualquier forma de gobierno la inteligencia puede realizar políticas en un amplio espectro que va desde el conservadurismo hasta el liberalismo, caracterizándose el primero por preferir la aplicación de soluciones ya conocidas, y el segundo por arriesgarse a elaborar soluciones nuevas. En definitiva, dice Sternberg, así como no existe un criterio único para evaluar la calidad de los gobiernos, tampoco existe un criterio universal para evaluar la calidad de las inteligencias. Una inteligencia anárquica y liberal, por ejemplo, puede ser exitosa para un músico y en cambio nefasta para un hombre de negocios.

Pudín

La tercera metáfora de la inteligencia que nos ocupa es de John Horn. Observando las definiciones de la inteligencia elaboradas durante los últimos ochenta años, Horn concluye que una buena metáfora para entender la inteligencia es la de un pudín (Horn, 1986, p. 92). La inteligencia, al igual que un pudín, no es unitaria, sino que está compuesta de muchas partes, y además esas partes en su opinión no forman un sistema, pues son altamente independientes entre sí y no interaccionan para producir fenómenos emergentes. Respecto a la definición de las partes, en el caso del pudín

una receta puede incluir uvas pasas, mientras que otra puede excluirlas, y sin embargo ambas son recetas válidas. Horn cree que con la inteligencia sucede lo mismo, y en consecuencia no tiene sentido seguir buscando la receta del pudín ideal, sino que la investigación debe centrarse en el examen pormenorizado de cada uno de los componentes posibles. El inconveniente de este enfoque, como él mismo reconoce, es que entonces se pierde la figura amplia de la capacidad intelectual humana, pero aún así está convencido de que es la mejor estrategia disponible en la actualidad. El análisis de Horn de las habilidades elementales que componen la inteligencia se basa en la teoría de la inteligencia fluida y cristalizada de su maestro Raymond Cattell, la cual a su vez está basada en la del mencionado Charles Spearman.

En la primera mitad del siglo XX el inglés Charles Spearman y el norteamericano Louis Thurstone ocupaban la escena central de las investigaciones en psicometría, la parte de la psicología dedicada a la medición de la mente. Mientras que Thurstone distinguía en la inteligencia siete factores primarios, Spearman sostenía una teoría bifactorial en la que sólo había uno de carácter primario, el famoso factor general *g*, siendo el resto de orden secundario o específico. Posteriormente, ya a mediados de siglo, Raymond Cattell dividió el factor general de Spearman en dos factores amplios e independientes: la inteligencia fluida (G*f*) y la inteligencia cristalizada (G*c*) (Davidson & Kemp, 2011, p. 60). La inteligencia fluida está relacionada con el procesamiento mental de la información novedosa y depende de la eficiencia del funcionamiento del sistema nervioso central, mientras que la inteligencia cristalizada consiste en el conjunto de habilidades e información que el sujeto adquiere y retiene en la memoria a lo largo de su vida. Ambas corresponden a la distinción de Donald Hebb entre inteligencia A e inteligencia B, en la cual A es la inteligencia que depende de factores exclusivamente biológicos al igual que la inteligencia fluida, y B es la que depende de la experiencia al igual que la inteligencia cristalizada (Eysenck, 1986, p. 69). Horn amplió la teoría de su maestro Cattell añadiendo en principio siete factores de primer orden a la inteligencia fluida y cristalizada, haciendo por tanto un total de nueve, e introdujo también un estrato inferior con más de ochenta factores de segundo orden en representación de

habilidades específicas asociadas a los factores de primer orden. La teoría conjunta de Cattell y Horn, denominada *teoría extendida de la inteligencia fluida y cristalizada*, sería más tarde remodelada por John Carroll para dar lugar a la *teoría CHC*, una de las más influyentes en la psicometría actual.

Así pues, los nueve factores de primer orden distinguidos por Horn serían metafóricamente los tropezones grandes del pudín, al tiempo que los de segundo orden serían los ingredientes de dichos tropezones, también variables según la receta. Es un modelo que recuerda al modelo atómico del pudín de uvas pasas propuesto por el físico Joseph Thomson, el descubridor del electrón. Thomson propuso que los electrones estaban distribuidos uniformemente alrededor del átomo de manera semejante a como se distribuyen las uvas pasas en un pudín. Se trataba de un modelo impreciso, que posteriormente fue sustituido por el modelo de Ernst Rutherford, mucho más informativo acerca de la distribución de las partículas subatómicas, ya que concentraba la masa en el núcleo y colocaba a los electrones orbitando en un cierto orden. De manera similar, la metáfora de Horn describe los elementos de la inteligencia pero no aborda su disposición sistemática, sino que niega la sistematicidad por principio, a pesar de que la observación y la experimentación neurofisiológica avalan la tesis de que la inteligencia funciona como un sistema modular, y no como un mero agregado de elementos independientes. La metáfora de Horn necesita ser perfeccionada así como el modelo de Rutherford perfeccionó el de Thomson. Quizás podría ser combinada con la analogía de Sternberg, la cual se ocupa de la jerarquía de los elementos pero no de su identificación exacta.

Calificación de una universidad

La cuarta metáfora que nos interesa es de Douglas Detterman, quien compara las formas de medir la inteligencia con las formas de medir la calidad de una universidad. Detterman comienza su exposición criticando la definición operacional de la inteligencia, según la cual la inteligencia sería el conjunto de medidas que predicen

el éxito académico (Detterman, 1986, p. 57). Se trata de una definición desgraciadamente muy extendida, en tanto que es consecuencia de la manera de pensar del positivismo, y el positivismo predomina en nuestra época. Como ya señalamos en el capítulo tercero, en términos operacionales el referente de un término consiste única y exclusivamente en el procedimiento para obtenerlo o medirlo (Marcuse, 1964, p. 38). Desde un punto de vista radicalmente opuesto, Detterman propone que la inteligencia puede ser definida como un conjunto finito de habilidades independientes entre sí que operan como un sistema complejo. Cuanto más complejo es un sistema, mayor es su grado de *totalidad* (*wholeness*), siendo éste muy alto en el caso de la inteligencia.

La inteligencia puede ser evaluada con una medida única como el factor *g* de Spearman, dice Detterman, de la misma manera que se puede puntuar la calidad de una universidad con una sola cifra. Este tipo de calificación sirve para discriminar a los mejores y los peores en cuanto a su rendimiento global, pero no sirve para averiguar cuáles son los aspectos más deficientes, y por tanto no tiene utilidad de cara a la intervención para mejorar los puntos débiles. La calificación global de una universidad depende de factores específicos de orden inferior tales como el nivel de excelencia del profesorado o el tamaño de la biblioteca, pero no informa de cuáles de ellos están rebajando la calificación global al hacer la media. Análogamente, los tests de inteligencia de factor general *g* son útiles para la discriminación de los individuos, pero no para identificar cuáles son los aspectos de la inteligencia que necesitan intervención terapéutica para mejorar el rendimiento global cuantificado por *g*.

En este punto se enfrentan dos grandes enfoques de la psicometría que ya hemos mencionado: el de factor general y el multifactorial. Los tests de inteligencia diseñados para obtener un factor general como el *g* de Spearman sólo sirven para la gradación de los sujetos en una escala, mientras que los tests que evalúan varios factores tienen la utilidad añadida de que proporcionan información para ayudar a los sujetos a mejorar sus competencias intelectuales. En la Historia de la psicometría el precursor tanto de Spearman como de Thurstone fue el francés Alfred Binet, quien en

su época, finales del siglo XIX, compitió con las tesis del británico Francis Galton, cuñado de Charles Darwin. Se considera a ambos, Galton y Binet, como los fundadores de la psicometría, aunque desde planteamientos totalmente distintos. Mientras que Galton identificaba la inteligencia con las capacidades psicofísicas de discriminación sensorial de estímulos, Binet consideraba que la inteligencia era un atributo más intelectual, dependiente de las habilidades complejas del juicio (Sternberg, 1999, p. 409). Según Binet, las habilidades fundamentales implicadas en la inteligencia son tres: dirección, adaptación y control. La dirección es el conocimiento de lo que se debe hacer y cómo debe hacerse, la adaptación selecciona y monitoriza las estrategias empleadas, y el control es la capacidad para criticar los propios pensamientos.

Con el paso del tiempo el enfoque psicofísico de Galton fue perdiendo adeptos y en los tests de inteligencia prevaleció el enfoque más intelectual de Binet, pero no así las intenciones que él tenía. El propósito del psicólogo francés era evaluar los citados tres elementos de la inteligencia para ayudar a los escolares de su país a mejorar su rendimiento académico (García, 2001a, p. 83), una noble intención que se vio malograda un par de décadas después con la aparición del primer test de inteligencia de factor general ideado por Spearman. La polémica en torno a la psicometría reside justo aquí. Cuanto más abstracto es un concepto, más diferencias elimina, y el factor general *g* las elimina todas, dado que reduce la medición de la inteligencia a una sola cifra. Como sostiene Detterman con su metáfora, los tests de inteligencia deben ser multifactoriales, pues de lo contrario se convierten en concreciones extremas del operacionalismo. El famoso *CI*, que erróneamente se suele denominar "coeficiente intelectual" cuando en realidad es el acrónimo de *cociente intelectual*, es una medida de la inteligencia que por sí sola es irreal e inhumana (Gardner, 1993, p. 241). Lo que significa el CI es la medida sobre 100 de la inteligencia general de un sujeto relativa a un conjunto de competencias y a un segmento de la población, siendo 100 la puntuación media obtenida por una muestra representativa de los sujetos pertenecientes a dicho segmento. Allí donde se utiliza el CI como baremo intelectual supremo, se está tratando a los seres humanos como objetos.

Habilidad atlética

La quinta y última metáfora que vamos a examinar es la de Robert Glaser, que compara la inteligencia con una habilidad atlética (Glaser, 1986, p. 77). Es una analogía esclarecedora que, a nuestro juicio, debe su éxito al hecho de que toda habilidad atlética es expresión de por lo menos una inteligencia: la cinético-corporal distinguida por Gardner. Según Glaser, la competencia intelectual (*intellectual proficiency*) comparte con la competencia atlética once características: una estructura de conocimiento disponible sobre el dominio de la competencia, componentes de realización automatizados, división de los eventos en fragmentos que permiten el reconocimiento de patrones, factores fisiológicos hereditarios, una disminución de la calidad de la ejecución causada por la falta de práctica y el paso de los años, un entendimiento tácito por parte de los expertos que en ocasiones les impide verbalizar con precisión las particularidades de su competencia, diferencias individuales que pueden ser denominadas como estilos, una maestría específica que no tiene por qué ser utilizable en otros dominios, mejora del rendimiento gracias a la utilización de artefactos tecnológicos y equipamiento en general, influencia de las condiciones motivacionales y, por último, la posibilidad de innovar en ese ámbito cuando se alcanza el más alto grado de maestría.

Dentro de las competencias intelectuales humanas Glaser traza una interesante distinción entre las artefactuales y las naturales. Las *artefactuales* son aquellas que se adquieren sobre todo en la escuela, mientras que las *naturales* se refieren a las competencias que se adquieren de manera espontánea sin educación reglada. Ejemplos de estas últimas son el conocimiento ontológico, el lenguaje natural, las habilidades matemáticas básicas y, en general, la capacidad de captar regularidades. Las competencias artefactuales son todas las demás, es decir, las adquiridas en el entorno académico, tales como la física, la química, la biología, la economía, la geografía, la Historia, y toda forma elaborada de conocimiento. Decimos que esta

distinción de Glaser es interesante porque traza justamente la línea divisoria de las competencias que las inteligencias artificiales han mostrado hasta la fecha ser capaces e incapaces de adquirir. Los sistemas expertos, a los que ya nos hemos referido anteriormente como uno de los mayores logros de la IA, son en algunos casos más hábiles que cualquier ser humano en el ámbito de las competencias artefactuales. Así, ningún ser humano juega al ajedrez mejor que el mejor de los sistemas expertos ajedrecistas, o ningún ser humano calcula tan rápido como una calculadora.

Pero cuando cambiamos de campo y nos situamos en el de las competencias naturales, encontramos que un niño normal de cinco años se desenvuelve en el mundo físico y domina el mundo social a través del lenguaje natural mejor que cualquier computadora electrónica. Es cierto que la computadora puede realizar minuciosos análisis sintácticos y morfológicos de las oraciones, dado que se trata de competencias artefactuales contenidas en los libros de gramática, pero en el aspecto natural de la competencia lingüística el niño es muy superior. La razón de esta asimetría radica en que las inteligencias artificiales no son capaces de adquirir competencias naturales, ya sean físicas o sociales, porque se trata de procesos basados en el *círculo hermenéutico*, entendido éste en el sentido amplio de la relación bidireccional de codeterminación entre el todo y las partes, tanto del texto como de cualquier otro signo. Para que una IA adquiriese competencias naturales sería necesario formalizar algorítmicamente el círculo hermenéutico, algo que, como veremos, es posible a nivel subsimbólico pero no al simbólico. A este asunto le dedicaremos especial atención en las siguientes secciones, exponiendo el funcionamiento del círculo hermenéutico a nivel representacional y neuronal, es decir, simbólico y subsimbólico.

5.2. Teorías de la inteligencia

Son tres las teorías contemporáneas de la inteligencia que vamos a examinar, en concreto las de Jeff Hawkins, Roger Schank y Howard Gardner. Antes de abordarlas es preciso que las ubiquemos en un esquema lo más exhaustivo posible de la

inteligencia, que tomaremos de Robert Sternberg. Según Sternberg la inteligencia puede investigarse en tres lugares, que son el individuo (I), el ambiente (II) y la interacción entre el individuo y el ambiente (III) (Sternberg, 1986, p. 4). A su vez, la inteligencia en el individuo puede ser analizada en tres niveles, que son el biológico (I.A), el molar (I.B) y el conductual (I.C). La teoría de Jeff Hawkins corresponde al nivel biológico (I.A) en tanto que describe la herramienta computacional común a todas las redes de neuronas corticales, la de Roger Schank se sitúa en el nivel molar (I.B) dado que examina la cognición en sentido representacional, y la de Howard Gardner se puede inscribir en el nivel conductual (I.C) si atendemos a que las inteligencias múltiples se refieren a diversos dominios de la conducta.

Dentro del nivel biológico (I.A) se puede estudiar la inteligencia a través de los organismos (I.A.1), dentro de los organismos (I.A.2) o bien en la interacción de los factores biológicos que operan a través y dentro de los organismos (I.A.3). A través de los organismos la inteligencia puede ser observada a lo largo de la evolución de las especies (I.A.1.a), en la genética de una especie en particular (I.A.1.b) o en la interacción que existe entre la evolución de las especies y la genética de determinadas especies en particular (I.A.1.c). Por su parte, dentro de los organismos la inteligencia puede ser examinada en términos de estructuras anatómicas (I.A.2.a), procesos fisiológicos (I.A.2.b) o de interacción entre dichas estructuras y procesos (I.A.2.c). Y finalmente, el análisis de la interacción de los factores biológicos que operan a través y dentro de los organismos ha de ser señalado como el más completo de los tres, ya que puede incluir todos los detalles de los otros.

El nivel molar (I.B) teoriza sobre dos aspectos del funcionamiento mental, que son la cognición (I.B.1) y la motivación (I.B.2). Dentro de la cognición algunos expertos distinguen entre la metacognición (I.B.1.a) y la cognición ordinaria (I.B.1.b). La metacognición se divide en procesos (I.B.1.a.i), conocimientos (I.B.1.a.ii) e interacción entre procesos y conocimientos (I.B.1.a.iii). De manera análoga, la cognición ordinaria, o simplemente cognición, también se divide en procesos (I.B.1.b.i), conocimientos (I.B.1.b.ii) e interacción entre procesos y conocimientos (I.B.1.b.iii). La diferencia entre

la metacognición y la cognición estriba que la *metacognición* se refiere a los procesos y los conocimientos sobre la propia cognición, mientras que la *cognición ordinaria* se refiere a lo que es conocido y controlado por la metacognición. Sternberg lo aclara con un par de ejemplos (Ibíd., p. 5). La metacognición como conocimiento, dice, sería el ser consciente de lo que uno sabe y no sabe, pues no es lo mismo saber o no saber algo que además saber o no saber que se sabe o no se sabe, mientras que la cognición como conocimiento sería el conocimiento en sí mismo. Por el otro lado, la metacognición como proceso de control consistiría en la elaboración de una estrategia para solucionar un problema. En cambio, la cognición como proceso controlado serían los pasos mentales usados para resolver el problema.

Las tres habilidades fundamentales implicadas en la inteligencia según Binet, dirección, adaptación y control, serían claros ejemplos de habilidades metacognitivas (Sternberg, 1999, p. 409), pues todas ellas implican la reflexión de la conciencia sobre sí misma. También Marvin Minsky recoge la distinción entre cognición ordinaria y metacognición en su *teoría de la sociedad de la mente*, en la que propone la existencia de un cerebro A y un cerebro B (Minsky, 1985, p. 59). El cerebro A sería controlado por el cerebro B, de manera que el primero actuaría como agente cognitivo y el segundo como agente metacognitivo. Dentro de la metacognición, los procesos pueden ser de planificación, supervisión o evaluación, mientras que los conocimientos pueden ser sobre la persona, la tarea–estrategia, o el contexto (García, 2001b, p. 10).

Continuando con el esquema de Sternberg, dentro de los procesos de la cognición ordinaria (I.B.1.b.i) se distingue entre la atención selectiva (I.B.1.b.i.a), el aprendizaje (I.B.1.b.i.b), el razonamiento (I.B.1.b.i.c), la resolución de problemas (I.B.1.b.i.d) y la toma de decisiones (I.B.1.b.i.e). Finalmente, en cuanto al nivel motivacional (I.B.2), se puede diferenciar entre el nivel o magnitud de la energía (I.B.2.a), la dirección o disposición de la energía (I.B.2.b) y la interacción entre el nivel y la dirección (I.B.2.c). Como ya dijimos al exponer la estructura y el funcionamiento del sistema nervioso, la información procedente del sistema motivacional participa en la formación de las representaciones mentales en las áreas de asociación. Así se entiende

que la motivación suponga a veces la diferencia entre el éxito y el fracaso hasta en las tareas cognitivas más sencillas como, por ejemplo, coger una pelota al vuelo (Kandel, Schwartz & Jessell, 1995, p. 613). La motivación puede ser intrínseca o extrínseca (García, 2001b, p. 11). La motivación intrínseca es la propia de aquellas personas que se dedican a una tarea por puro placer y autosatisfacción, mientras que la extrínseca busca recompensas proporcionadas por el mundo exterior, tales como el reconocimiento, el prestigio social o el dinero.

Por último, respecto al nivel conductual (I.C) Sternberg distingue entre los dominios académico (I.C.1), social (I.C.2) y práctico (I.C.3). El académico abarcaría las competencias que Robert Glaser denomina artefactuales, es decir, aquellas que se adquieren en la escuela y otras instituciones de formación reglada, mientras que los otros dos, el social y el práctico, corresponderían a las competencias naturales. Acerca del nivel académico (I.C.1), señala Sternberg, existen dos grandes controversias. La primera es la amplitud de lo que debe caer bajo la extensión de la inteligencia, ya que si todo dominio de la conducta humana es calificable de inteligente entonces ésta termina siendo un concepto que pierde su significado. La segunda es acerca de la especificidad de los dominios, pues algunos autores los trazan amplios mientras que otros los definen estrechos y en consecuencia acaban distinguiendo un número muy elevado. La mayoría, dice Sternberg, está de acuerdo en que debe haber dominios generales (I.C.1.a), dominios específicos (I.C.1.b) y una interacción entre los dominios generales y específicos (I.C.1.c). Por su parte, en el dominio social (I.C.2) cabe diferenciar entre la inteligencia aplicada al autoconocimiento (I.C.2.a), que sería el equivalente a la inteligencia intrapersonal en la teoría de Gardner, la aplicada al conocimiento de los otros (I.C.2.b), que equivaldría a la inteligencia interpersonal, y la interacción entre ambas (I.C.2.c). El tercer gran dominio de la conducta es el de lo práctico (I.C.3), que incluye aspectos ocupacionales (I.C.3.a), cotidianos (I.C.3.b) y de interacción entre los dos (I.C.3.c). Los aspectos ocupacionales, dice Sternberg, incluyen el saber cómo realizar el trabajo de manera eficaz y cómo adelantar tarea. Los cotidianos se refieren a acciones ordinarias.

Antes de exponer las teorías de Jeff Hawkins, Roger Schank y Howard Gardner para explicar la inteligencia respectivamente a nivel cerebral (I.A), mental (I.B) y conductual (I.C) debemos justificar su elección respecto de otras alternativas como, por ejemplo, las teorías de Hans Eysenck (I.A), Robert Sternberg (I.B) y Robert Glaser (I.C). La elección de Hawkins y Schank se debe, en primer lugar, a que plantean enfoques actualizados de la neurociencia y la psicología cognitiva respectivamente, que son las dos disciplinas en las que se basan la IA subsimbólica y la IA simbólica. Segundo, porque conocen bien las dificultades técnicas de la IA dado que ambos han dedicado buena parte de sus carreras profesionales a la programación informática. Y tercero, porque se centran en un problema que, a nuestro juicio y el de Hubert Dreyfus, ha supuesto una de las mayores dificultades para el avance de la IA: la estructura circular de la comprensión (Dreyfus, 1992, p. 60). El círculo hermenéutico es una de las claves para la creación de verdaderas inteligencias artificiales en sentido fuerte, esto es, con competencias naturales, y es un asunto central en las teorías de Hawkins y Schank. Ninguno de los dos lo menciona explícitamente por su nombre, pero en las lecturas que haremos de sus textos se podrá apreciar que es un tema recurrente. En cuanto a la elección de Gardner, ya hemos dicho que nos parece la teoría contemporánea de la inteligencia más acertada, y por eso la hacemos nuestra, debido a su carácter integral que estudia la inteligencia vinculando el nivel mental con los niveles biológico y cultural. Su enfoque holista, propio de un humanista como es él, responde a la realidad compleja de la inteligencia, a diferencia de los planteamientos eliminativistas del fisicalismo y el cognitivismo.

5.2.1. La inteligencia en el cerebro

De acuerdo a la *hipótesis de Mountcastle*, toda la corteza cerebral funciona utilizando una sola herramienta computacional. Jeff Hawkins adopta este supuesto como fundamento de su teoría, y en consecuencia afirma que la corteza cerebral es una memoria autoasociativa cuya herramienta computacional común consiste en

elaborar representaciones invariables y utilizarlas para realizar predicciones (Hawkins & Blakeslee, 2004, p. 66), entendiendo el predecir en un sentido amplio que abarca tanto el *anticipar* lo que va a suceder en el futuro como el *completar* el presente con información no proporcionada por las entradas sensoriales. Dado que la validez de la teoría de Hawkins depende por completo de la hipótesis de Mountcastle, debemos comenzar examinando esta última. Hay numerosas pruebas en favor de la hipótesis de Mountcastle, siendo la más contundente la plasticidad de la corteza cerebral.

Como ya apuntamos en el capítulo anterior, es un hecho demostrado que cuando una persona que no es ciega de nacimiento adquiere una ceguera total, al cabo de un tiempo su corteza visual, ubicada en el lóbulo occipital, cambia su función y pasa a dedicarse al procesamiento de información táctil para leer en braille. Un caso aún más sorprendente es el del aparato de visualización diseñado por Paul Bach y Rita (Ibíd., p. 78), que consiste en una pequeña cámara de vídeo acoplada en la frente de un sujeto que se ha quedado ciego y en un dispositivo reticular de presión que se le coloca en la lengua. Cada píxel de la imagen captada por la cámara está asociado ordenadamente a un punto de presión del dispositivo reticular, de manera que se conserva la organización topográfica. Cuanto mayor es la intensidad lumínica de un píxel, más profundamente se hunde en la lengua su punto de presión asociado. El resultado parece de ciencia ficción, pero es real, y es que el sujeto termina viendo el mundo a través de la lengua. Los casos de este tipo evidencian una plasticidad neuronal tan versátil que sólo es posible si, como propone Mountcastle, el algoritmo básico de las redes de neuronas es el mismo para toda la corteza.

No obstante, Hawkins reconoce que la correspondencia de las funciones con ciertas zonas de la corteza es también un hecho innegable. Así, por ejemplo, aunque el algoritmo sea común, no deja de ser cierto que las regiones corticales dedicadas al lenguaje están ubicadas en casi todos los individuos en el hemisferio izquierdo, concretamente en las zonas de Broca, Wernicke y algunas más, todas en las inmediaciones del lóbulo temporal. Pero esta especificidad determinada genéticamente, dice Hawkins, no es incompatible con la hipótesis de Mountcastle, sino

sólo una prueba de que la evolución ha encontrado una división funcional de la corteza que resulta lo bastante eficaz como para mantenerse en el genotipo humano. En caso de lesión cortical, la plasticidad permite al cerebro encontrar otras divisiones funcionales alternativas.

Otro dato científico en favor de la hipótesis de Mountcastle es que los potenciales de acción disparados por las neuronas no contienen la modalidad o submodalidad de la información transmitida. Como ya dijimos en el capítulo anterior, el principio organizativo del procesamiento en paralelo establece que los sistemas sensorial, motor y motivacional están compuestos de subsistemas que son independientes a nivel anatómico y funcional (Kandel, Schwartz & Jessell, 1995, p. 84). Debido a esta división, las cuatro submodalidades del sistema somatosensorial transmiten su información por vías paralelas. El tacto, la propiocepción, el dolor y la temperatura recorren nervios diferentes, de manera que el cerebro sabe que una sensación es de dolor y no de temperatura porque le llega a través de la vía del dolor. Si se cruzaran, se percibiría el dolor como temperatura y la temperatura como dolor. Esto es lo que les sucede a las personas con *sinestesia*, una enfermedad que habitualmente da lugar a la visión de los sonidos o la audición de las imágenes, aunque técnicamente puede presentarse como la combinación de cualesquiera modalidades sensoriales. Los potenciales de acción que transmiten la información auditiva no contienen ningún indicador que permita al cerebro saber que se trata, en efecto, de información auditiva y no visual. Esto avala la hipótesis de Mountcastle, en tanto que la corteza se dedicaría sólo a procesar patrones de disparo de potenciales de acción sin importar el lugar del que procedan o al que se dirijan. «Las entradas al cerebro son sólo patrones. No importa de dónde provienen éstos; siempre que tengan una correlación temporal coherente, el cerebro puede hallarles sentido. [...] A la corteza cerebral no le importa si los patrones se originaron en la visión, el oído u otro sentido. No le importa si sus entradas provienen de un único órgano sensorial o de cuatro» (Hawkins & Blakeslee, 2004, p. 78). El cerebro es una caja oscura que sólo recibe patrones eléctricos con los que construye representaciones de lo que hay ahí fuera.

Ya hemos mencionado que el cerebro, según Hawkins, no es un procesador de información similar a una computadora electrónica, sino un *sistema de memoria*. A nivel estructural la diferencia estriba en que en la computadora, tal y como establece la arquitectura von Neumann, el núcleo de procesamiento y el almacén de memoria son estructuras distintas, mientras que en un sistema de memoria como el cerebro ambas tareas, el *procesamiento* y la *memorización*, se realizan en la misma estructura: la neurona. A nivel funcional las diferencias se concretan en cuatro atributos exclusivos de la memoria cortical que no están presentes en las memorias informáticas. Estos cuatro atributos son que la corteza cerebral almacena secuencias de patrones, recuerda los patrones por autoasociación, los almacena en una forma invariable y los estructura en una jerarquía (Ibíd., p. 88).

Secuencias de patrones

Respecto al primer atributo, Hawkins propone un sencillo experimento mental para mostrar que la corteza cerebral *almacena secuencias de patrones*, y no fragmentos aislados de información. Todos sabemos recitar el alfabeto de la A a la Z con facilidad, dice, pero si nos piden que lo hagamos al revés, de la Z a la A, entonces la tarea se torna muy complicada. En cambio, una computadora electrónica es capaz de devolver una secuencia de elementos en el orden en el que le fueron introducidos o en el inverso sin que ello suponga una diferencia apreciable en el tiempo de respuesta. Esta desemejanza se debe a que la memoria de la computadora electrónica registra los elementos desligados, sin ningún vínculo entre ellos, mientras que el cerebro humano almacena los patrones de información en forma de secuencias.

El caso del alfabeto es el de una memoria explícita, pero el mismo fenómeno se observa también en las memorias implícitas. Atendiendo a su contenido las memorias pueden ser explícitas o implícitas (Kandel, Schwartz & Jessell, 1995, p. 656). Las *memorias explícitas* o *declarativas* contienen los conocimientos conscientes de *qué* son las cosas, ya sean sucesos particulares codificados en forma de *memorias*

episódicas o generalizaciones codificadas en *memorias semánticas*. Por su parte, las *memorias implícitas* o *procedimentales* contienen los conocimientos inconscientes de *cómo* es el mundo y cómo habérnoslas con él, y pueden ser *asociativas* o *no asociativas*, en función de si asocian elementos distintos. El *condicionamiento clásico* y el *condicionamiento operante*, vistos en el capítulo anterior, son los dos grandes tipos de memorias implícitas asociativas, pues el primero registra la asociación de dos estímulos diferentes, y el segundo, de una conducta y un estímulo. En cuanto a las memorias implícitas no asociativas, las más importantes son la habituación y la sensibilización. La *habituación* consiste en la repetición del mismo estímulo de manera que el sujeto termina percibiéndolo con menor intensidad, mientras que en la *sensibilización* el resultado es el contrario y la intensidad percibida aumenta.

Un ejemplo de memoria implícita o procedimental sería aquella que utilizamos para conducir un automóvil. Ciertamente, cuando damos nuestra primera clase práctica en la autoescuela nuestro conocimiento sobre cómo manejar el coche es todavía meramente explícito o declarativo. Durante el estudio de la parte teórica para obtener el carné hemos memorizado un conjunto de proposiciones que abarcan desde las normativas de seguridad vial hasta cómo cambiar de marcha. A partir de la primera clase práctica se inicia el proceso para convertir todas esas memorias declarativas en memorias procedimentales (García, 2009, p. 16). Al principio tenemos que focalizar nuestra atención en las proposiciones aprendidas durante la parte teórica, pero con el paso del tiempo vamos automatizando su ejecución hasta que, por fin, conducir termina siendo una tarea casi tan inconsciente como el caminar. Ahora bien, si hemos aprendido a conducir con un coche de cambio manual, entonces la primera vez que nos pongamos al volante de uno automático lo normal es que al escuchar que suben las revoluciones del motor busquemos sin pensarlo el pedal de embrague y la palanca de cambios, porque ésa es la secuencia a la que estamos acostumbrados: cuando suben las revoluciones, piso el embrague y muevo la palanca para cambiar de marcha. Este caso ilustra que las memorias implícitas, al igual que las explícitas, también son almacenadas en el cerebro en forma de secuencias de patrones.

Autoasociación

El segundo rasgo característico de la memoria cortical recogido por Hawkins es que la corteza *recuerda los patrones por autoasociación*. Esto quiere decir que los patrones están asociados consigo mismos, de manera que la presencia de una parte es suficiente para recuperar el resto. Por ejemplo, con sólo ver los ojos de una persona basta para que sepamos que es ella, pues ese fragmento de información evoca todos los demás, desde su nombre hasta el resto de su cara o el olor de su pelo. Esto sucede gracias al principio de Hebb, el cual establece, recordemos, que cuando una neurona A participa repetida o persistentemente en la excitación o inhibición de otra neurona B, entonces acontece algún tipo de proceso o cambio metabólico en una o ambas que incrementa la eficacia de A para excitar o inhibir a B (Fuster, 1997, p. 681). El principio de Hebb suele resumirse en inglés con un juego de palabras: *neurons that fire together, wire together* (las neuronas que disparan juntas, se conectan juntas). Dado que los ojos suelen presentarse junto con el resto de la cara de una persona, las redes neuronales implicadas en el reconocimiento de esos ojos, de esa boca, y demás partes de ese rostro suelen activarse al mismo tiempo, y en consecuencia la activación de una sola de ellas da lugar a la activación del resto. «En cualquier momento una parte puede activar el todo. Esta es la esencia de las memorias autoasociativas» (Hawkins & Blakeslee, 2004, p. 93). Y éste es también el mecanismo que opera cuando mantenemos una conversación en un ambiente ruidoso que nos impide escuchar todas las palabras. El cerebro completa las lagunas de información con suposiciones coherentes, igual que completa un rostro familiar con sólo ver una parte.

Sobre este asunto los experimentos de Michael Gazzaniga con pacientes comisurotomizados son reveladores (García, 2001a, p. 152). La comisurotomía consiste en la sección del cuerpo calloso, que es la principal vía de comunicación entre los hemisferios cerebrales. Es una intervención quirúrgica muy agresiva que se realiza, por ejemplo, sobre pacientes epilépticos que tienen el foco de los ataques situado en un

hemisferio, para evitar así que su efecto se extienda por todo el cerebro. Al encontrarse los hemisferios desconectados entre sí, es posible transmitirle a uno de ellos una información que no llegue a ser recibida por el otro. De esta forma, se puede dar al hemisferio derecho la orden de levantarse y caminar sin que sea conocida por el izquierdo. Dado que el módulo intérprete del lenguaje suele encontrarse en el izquierdo, cuando se le pregunta al paciente por qué se ha levantado y adónde va, enseguida elabora una respuesta verbal coherente con la situación, como por ejemplo "voy a por un vaso de agua porque tengo sed". El hemisferio izquierdo sólo dispone de la información de que está de pie y caminando, y con ese fragmento le basta para crear una explicación de la totalidad, un fenómeno que es posible gracias al carácter autoasociativo de la memoria.

Representaciones invariables

La tercera propiedad distintiva de la memoria cortical es que *almacena los patrones en forma de representaciones invariables*. Aquí reside el antes mencionado problema filosófico de los universales, que es abordado por Hawkins desde un punto de vista científico. Ya en el siglo IV a.C. Platón se preguntaba cómo es posible que seamos capaces de reconocer un objeto particular como perteneciente a una clase universal, cuando en la realidad ordinaria del mundo sensible no existen dos objetos particulares que sean idénticos entre sí. Y aunque fuesen idénticos, nuestra percepción de ellos no lo sería, pues es inevitable observarlos siempre desde perspectivas y bajo condiciones ambientales diferentes, nunca antes experimentadas. La solución de Platón a este problema epistemológico fue su célebre *teoría de la reminiscencia*. Tal y como la expone en el *Fedro*, cuando un hombre muere, la parte racional de su alma se separa de los apetitos irascible y concupiscible y asciende al cielo, desde donde contempla, en un viaje orbital guiado por Zeus, el lugar en el que se encuentran las Ideas universales de todas las cosas. «Es en dicho lugar donde reside esa realidad carente de color, de forma, impalpable y visible únicamente para el piloto del alma, el

entendimiento; esa realidad que "es" de una manera real, y constituye el verdadero objeto del conocimiento» (Platón, *Fedro*, 247c). Tras un cierto tiempo en la región donde habita el linaje de los dioses, el alma vuelve al mundo terrenal para reencarnarse. La capacidad entonces de reconocer los objetos particulares como pertenecientes a una clase universal, dice Platón, se debe a la semejanza de éstos con la Idea prototípica de dicha clase grabada en el alma. De esta manera, reconocemos que un objeto particular es una casa en la medida en que las impresiones sensibles que de él recibimos nos recuerdan a la Idea universal de casa.

Hawkins está de acuerdo en que el reconocimiento de los objetos particulares se debe a la semejanza de éstos con el equivalente a las Ideas universales en su terminología, que son las representaciones invariables, pero obviamente no comparte la teoría platónica en lo referente a la naturaleza de estas últimas y al proceso mediante el cual se adquieren. Según Platón, las representaciones invariables tienen un contenido intensional eterno, inmutable y universal para todos los hombres. El objetivo de la ciencia, como dice en el pasaje que acabamos de citar, sería expresar lingüísticamente dicha intensión. Hawkins, en cambio, no está de acuerdo en que sean eternas, inmutables ni universales, sino que afirma todo lo contrario: son transitorias, mutables y exclusivas del sujeto que las elabora.

Organización jerárquica

Los procesos de elaboración de las representaciones invariables y de reconocimiento de los objetos particulares como pertenecientes a dichas representaciones se basan ambos en la *organización jerárquica* de la corteza cerebral, que es justamente el cuarto rasgo distintivo de la memoria cortical enumerado por Hawkins. Recordemos que una neurona sólo dispara potenciales de acción si la sumación temporal y espacial de los potenciales electrotónicos producidos por las neuronas que proyectan sobre ella sobrepasa el umbral de disparo. De esta manera se conforma una estructura jerárquica abstractiva. En el caso del subsistema visual

dedicado al reconocimiento de formas, los fotorreceptores de la retina proyectan hacia las células bipolares, que a su vez proyectan hacia las células del núcleo geniculado lateral del tálamo, que a su vez proyectan hacia las células esteladas de la capa 4Cβ de la corteza visual primaria V1, que a su vez proyectan hacia las células complejas de V1, y así sucesivamente hasta llegar a la corteza inferotemporal IT (Kandel, Schwartz & Jessell, 1995, p. 440). Este esquema tiene la forma de un árbol convergente hacia un número cada vez menor de ramas, produciéndose una abstracción en cada nudo. Por ejemplo, en V1 las primeras neuronas del proceso responden a estímulos circulares, mientras que las siguientes reciben las proyecciones de las anteriores y sólo responden a estímulos lineales con un grado de rotación y una posición específicos, y a su vez las siguientes reciben las proyecciones de las anteriores y sólo responden a estímulos lineales también con un grado de rotación específico pero con independencia de su posición dentro del campo visual. A medida que las señales ascienden en la jerarquía, las particularidades van perdiendo importancia. Puede haber, por ejemplo, un cambio en la posición de un estímulo lineal, y eso afectará a la actividad eléctrica de las neuronas situadas en el mencionado nivel que es sensible a la posición, pero la actividad de las del nivel siguiente no se verá alterada. O podemos observar una sucesión discontinua de puntos y sin embargo percibirlos todos ellos como formando una línea recta.

Por consiguiente, las representaciones invariables se forman en todos los niveles de la jerarquía cortical, y la única diferencia entre ellas es su grado de complejidad (Hawkins & Blakeslee, 2004, p. 147). En los niveles más bajos las representaciones invariables formadas son sencillas, como los puntos o las líneas en el caso del sistema visual, mientras que en los más altos son complejas. Así descrito puede parecer que lo que sucede en la corteza cerebral es un proceso trivial, en tanto que sólo se trata de ir aumentando el grado de abstracción de las representaciones invariables, pero en realidad se trata de un mecanismo muy complejo que todavía no se conoce con exactitud, y por tanto no se dispone de un algoritmo para que una computadora electrónica lo simule.

Gracias a las cuatro propiedades que acabamos de exponer, la corteza cerebral es capaz de crear un modelo del mundo. El mundo, dice Hawkins, tiene una estructura jerárquica análoga a la de la corteza. Todos los objetos que nos rodean está compuestos por otros objetos de nivel inferior que aparecen siempre juntos. «Cuando asignamos un nombre a algo, lo hacemos porque hay un conjunto de rasgos que van constantemente juntos. Un rostro es un rostro debido a que siempre aparecen juntos dos ojos, una nariz y una boca. Un ojo es un ojo porque siempre aparecen juntos una pupila, un iris, un párpado y demás» (Ibíd., p. 149). Hawkins denomina a esto *estructura nido*. «La corteza cerebral posee un algoritmo de aprendizaje inteligente que descubre y capta de forma natural cualquier estructura jerárquica que exista. Cuando falta dicha estructura, caemos en la confusión e incluso en el caos» (Ibíd., p. 150). La inteligencia definida metafóricamente como máquina del tiempo funciona evocando patrones de nivel superior a partir de los de nivel inferior. Así, cuando vemos unos ojos asomándose por una ventana, predecimos que debajo hay una nariz, una boca y un cuerpo entero.

En esto se basan muchos chistes: en violar las predicciones razonables. En la secuencia final de *Annie Hall* (1977) Woody Allen cuenta una anécdota que nos sirve para ilustrarlo. Dice que un señor va al psiquiatra para contarle que está preocupado porque su hermano se cree que es una gallina. El psiquiatra le responde que lo que debería hacer es encerrarlo en un manicomio, a lo que él responde: "no puedo; necesito los huevos". Al escuchar la primera parte, uno supone que el hombre que acude al psiquiatra es una persona sensata, ya que se trata de alguien capaz de reconocer que un ser humano que se cree una gallina está loco, pero su frase final viola la suposición realizada al principio. Roger Schank, cuya teoría de la inteligencia vamos a examinar en la sección siguiente, considera que algunos chistes son formidables tests de inteligencia. Ello se debe a que cuanto más inteligente es alguien, más predicciones aplica sobre los estímulos que recibe, y por tanto más sonadamente reirá al escuchar un chiste de este tipo y descubrir al final que sus predicciones eran incorrectas (Schank, 1986, p. 127).

Realimentación

La habilidad predictiva es posible gracias a que la jerarquía cortical funciona no sólo en sentido ascendente, sino también descendente, un hecho que, como denuncia Hawkins, ha sido tradicionalmente olvidado por los investigadores de la IA subsimbólica, quienes durante décadas se han obstinado en crear redes de neuronas artificiales sólo de sentido ascendente (Hawkins & Blakeslee, 2004, p. 40). De acuerdo al principio de polaridad dinámica descubierto por Ramón y Cajal, las neuronas sólo pueden transmitir señales eléctricas en un sentido. La dualidad de sentidos del flujo eléctrico se logra por tanto mediante circuitos de *realimentación* o *retroalimentación* (*feedback*) que consisten en proyecciones de las regiones superiores hacia las inferiores. De esta manera, cuando una región superior recibe de *abajo a arriba* (*bottom-up*) información de regiones inferiores, su respuesta es proyectar de *arriba a abajo* (*top-down*) de vuelta hacia a ellas para, en cumplimiento del principio de Hebb, activar aquellas que no se hayan activado pero que suelen activarse en esas circunstancias. Así funciona la autoasociación a nivel neuronal. Si la predicción o la suposición para completar el sentido de algo lanzada por la región superior resulta exitosa, entonces se refuerza, de lo contrario, se debilita. En este último caso, el patrón de entrada continúa propagándose hacia arriba en la jerarquía hasta encontrar una región que realice una suposición o predicción exitosa. Si no hay ninguna, entonces entran en escena las habilidades metacognitivas, y el sujeto ha de aplicar el razonamiento para hallar una manera de responder al patrón entrante (Ibíd., p. 157).

En resumen, la inteligencia tal y como la entiende Hawkins es la capacidad para predecir de manera exitosa, entendiendo la predicción en un sentido amplio que abarca tanto el anticipar el futuro como el completar el presente con información no recibida del exterior en ese momento. «La inteligencia se mide por la capacidad de recordar y predecir patrones del mundo, incluidos lenguaje, matemática, propiedades físicas de los objetos y situaciones sociales» (Ibíd., p. 117). «Conocer algo significa que

puedes realizar predicciones al respecto» (Ibíd., p. 125). Un dato sorprendente que revela la importancia de las predicciones en la cognición es que sólo entre un 10 y un 20% de las conexiones que proyectan sobre el núcleo geniculado lateral del tálamo provienen de la retina. El resto procede de otras regiones del cerebro y son conexiones de realimentación (Kandel, Schwartz & Jessell, 1995, p. 431). Por tanto podemos decir que de toda la información que llega a la corteza visual primaria, que es el destino preferente de las proyecciones del núcleo geniculado lateral, tan sólo un 15% aproximadamente procede de los ojos, y el resto son informaciones elaboradas por el cerebro. A lo cual habría que añadir las numerosas conexiones de realimentación intermedias que suceden con posterioridad, ya que la corteza visual primaria es, como su nombre indica, una región de bajo nivel muy alejada del resultado final de lo que vemos. Estos datos avalan la tesis comúnmente enunciada de que la mayoría de las veces vemos lo que esperamos ver y no lo que nuestros ojos ven.

A nuestro modo de ver, parafraseando la famosa metáfora utilizada por Frederick Copleston para explicar la teoría del conocimiento de Kant, la inteligencia puede ser descrita como una suerte de tecnología natural de gafas de *realidad aumentada* que muestra lo que está presente relacionándolo con algo ausente. En opinión de Hawkins, para que una máquina o un animal merezca el calificativo de inteligente no es necesario que se comporte en términos generales como un ser humano, sino sólo que sea capaz de realizar buenas predicciones en base a la memoria de lo ausente: «La inteligencia se mide por la capacidad predictiva de una memoria jerárquica, no por una conducta semejante a la humana» (Hawkins & Blakeslee, 2004, p. 242). A través de la metáfora de la inteligencia como realidad aumentada se aprecia que la inteligencia es un fenómeno similar a la locura. El loco y el genio son aquellos que ven lo que los demás no ven. En el chiste de Woody Allen, el protagonista del relato está loco porque cree que hay huevos donde en realidad no los hay, o al menos donde los demás estamos convencidos de que no es razonable creer que los hay. Por eso hemos escogido este chiste y no otro. El loco y el genio son considerados excéntricos porque tienen visiones que nadie más comparte. Newton veía fuerzas

donde los científicos de su tiempo veían impulsos aristotélicos. Kaspárov veía jugadas de ajedrez que sus adversarios no eran capaces de ver, y Deep Blue veía jugadas que ni siquiera Kaspárov era capaz de ver.

El único criterio para discriminar al genio del loco es el éxito de sus visiones, justo el mismo criterio que, desde el enfoque instrumental de la ciencia que venimos defendiendo, decide la selección de las teorías científicas. La coincidencia se debe a que las teorías científicas también son visiones de cosas que no se ven, como las fuerzas de la mecánica newtoniana, o la relación entre la masa y la energía. Las visiones de realidad aumentada del genio resultan ser exitosas, mientras que las del loco conducen al fracaso, siempre dentro de un contexto social que decide qué es el éxito y qué es el fracaso. El contexto social cambia con el tiempo, y así quien antes era un loco puede después ser un genio, como le ocurrió a Nietzsche. La visión de la realidad en sí misma, sin aumento subjetivo, sería inútil, si es que fuera posible.

Las nuevas gafas de Kant

Antes de pasar al examen de la inteligencia en la mente, detengámonos un momento en el aumento subjetivo de la realidad para distinguir sus partes. Por su ámbito de aplicación, los aumentos subjetivos de la realidad pueden ser relativos al *mundo físico* o al *mundo social*. Por su origen pueden ser *genéticos* o *ambientales*. Y por su alcance poblacional pueden ser *particulares* o *universales*. Las célebres gafas de Kant, por tanto, antes que parecerse a unas gafas comunes, se parecen más bien a esas monturas que utilizan los oftalmólogos en las pueden colocarse al mismo tiempo varias lentes superpuestas, cada una con un ámbito de aplicación, como reducir la miopía, la hipermetropía o el astigmatismo.

Las lentes genéticas serían las determinadas por factores heredados de manera biológica. Dentro de ellas, las particulares son las que diferencian a unos individuos de otros, como la tendencia a la depresión o la hiperactividad. Casos extremos serían las personas que perciben el mundo físico de manera peculiar debido a la sinestesia, o las

que tienen una percepción distorsionada del mundo social por culpa del autismo. En cambio, las universales son compartidas por toda la especie humana. A este último grupo pertenecerían las formas *a priori* de la sensibilidad y los conceptos puros del entendimiento descubiertos por Kant a finales del siglo XVIII. El error del gran filósofo alemán fue creer que esas estructuras debían su universalidad a un origen metafísico, mientras que hoy en día, gracias a la teoría de la evolución de Darwin, sabemos que se trata de rasgos heredados por selección natural (Habermas, 1968b, p. 174), de la misma manera que todas las cigüeñas tienen la habilidad innata de construir nidos. Lo que para nosotros son estructuras universales de la cognición *a priori*, es decir, innatas, desde el punto de vista de la especie humana son *a posteriori*, esto es, adquiridas por selección natural tras miles de años de adaptación (García, 2010, p. 10). Si la causalidad, que es un concepto puro del entendimiento en la teoría de Kant, no hubiera resultado útil para la supervivencia, entonces habría desaparecido, al igual que perdimos la cola cuando adoptamos la postura bípeda, pues ya no necesitábamos un apéndice para equilibrar el cuerpo. La estructura circular de la comprensión sería otro ejemplo de estructura genética universal de aumento subjetivo de la realidad, que afecta tanto a la comprensión del mundo físico como del mundo social.

En cuanto al aumento debido a factores ambientales, también puede ser particular o universal, si bien entendiendo la universalidad en un sentido restringido que no abarca a toda la especie humana, sino sólo a una comunidad. Cada individuo posee un juego de lentes particulares como resultado de su trayectoria vital única e irrepetible. Por ejemplo, don Quijote, dice Cervantes, era un hombre como todos los demás hasta que en un momento determinado «del poco dormir y del mucho leer, se le secó el cerebro», es decir, que adquirió una visión marcadamente singular de la realidad a consecuencia de una serie de experiencias personales, la mayoría vividas en su biblioteca. Respecto a las lentes universales, podemos denominarlas así en tanto que son las compartidas por una comunidad. Esto es lo que son todas las culturas: formas de ver la realidad que cada generación transmite a la siguiente a través de la tradición, ya sea oral o escrita. Un ejemplo claro son los proverbios. Entre otras

muchas cosas, en la siguiente sección vamos a ver que Roger Schank clasifica los proverbios dentro de un tipo de estructura memorística a la que él denomina TOP (*thematic organization packet*).

5.2.2. La inteligencia en la mente

Juguemos a las 20 preguntas. Intervienen dos participantes. Uno de ellos piensa en un personaje famoso, y el otro puede hacerle 20 preguntas como máximo para averiguar la identidad de dicho personaje. Supongamos que, tras haber agotado la mitad de las cuestiones, el interrogador tiene los datos de que se trata de un filósofo moderno. Con esta información los candidatos que pasan por su cabeza supongamos que son John Locke, David Hume, Gottfried Leibniz y Baruch Spinoza. El siguiente dato obtenido es que se trata de un racionalista. Lo que sucede entonces en la mente del interrogador es que comprueba cuáles de los candidatos cumplen con el nuevo atributo y cuáles no. De esta manera, caen de la lista Locke y Hume. El resto son racionalistas: Leibniz y Spinoza. Para afinar en la discriminación, al interrogador se le ocurre entonces preguntar cuántas sustancias componen la realidad según el filósofo en cuestión, y recibe la respuesta de que es dualista. Esto supone un problema, porque ni Leibniz ni Spinoza son dualistas. Tras pensarlo un poco, se le ocurre un nuevo candidato: René Descartes. Encaja a la perfección en el perfil provisional, ya que se trata de un filósofo moderno, racionalista y dualista. El interrogador se arriesga a preguntar si es Descartes, y, en efecto, lo es. Ha ganado el juego.

En términos neuronales lo que ha ocurrido en el cerebro del interrogador es que, tras obtener cada nueva información, se ha producido un *cruce* entre las representaciones *ascendentes* de bajo nivel y las *descendentes* de alto nivel. De abajo a arriba se proyectaban las pistas, y de arriba a abajo se proyectaban los candidatos. Cuando un candidato no se ajustaba a las pistas, era descartado. Así es como cayeron de la lista Locke y Hume cuando se supo que la pertenencia a la tradición racionalista era un atributo. Y en sentido inverso, las pistas sugerían a los candidatos. Es lo que

ocurrió cuando la lista de candidatos se quedó vacía porque ninguno de ellos sostenía un dualismo de sustancias. Las pistas sugirieron entonces el nombre de Descartes. Jeff Hawkins describe así el proceso: «La intersección de estas dos sinapsis (de arriba-abajo y de abajo-arriba) nos proporciona lo necesario. [...] Este mecanismo no solo realiza predicciones específicas, sino que también resuelve ambigüedades de las entradas sensoriales. [...] Este mecanismo de correspondencia de abajo-arriba/arriba-abajo nos permite decidir entre dos o más interpretaciones. [...] Donde los dos conjuntos se intersecan es lo que percibimos. [...] Así es como decidimos si la foto es de un jarrón o de dos caras» (se refiere a la ilusión óptica del jarrón de Rubin) (Ibíd., p. 180). De abajo a arriba las partes proyectan hacia el todo, y arriba a abajo el todo proyecta sobre las partes. Así funciona el *círculo hermenéutico*.

La diferencia entre el círculo hermenéutico y un círculo vicioso es que en este último los sucesivos giros son improductivos, mientras que en el círculo hermenéutico cada giro del pensamiento conduce a una comprensión cada vez más afinada. Al dar una nueva vuelta el interrogador descubrió que algunos candidatos debían ser descartados. Al ganar el juego, la representación de nivel superior, Descartes, iluminó la comprensión de las representaciones de nivel inferior de una manera distinta. Cuando se pensaba en un filósofo moderno al principio, tal vez se conjeturaba que sería inglés o alemán, pero al descubrir que la respuesta era Descartes, entonces Francia le vino a la mente. Las partes determinan el todo y el todo determina las partes. Así se van llenando los huecos y se va cerrando el *abismo* que separa al sujeto y al objeto de la comprensión. Además de la relación circular entre el *todo* y las *partes*, es importante advertir que en el círculo hermenéutico opera simultáneamente una relación entre lo *anterior* y lo *posterior*, no sólo porque las partes anteriores configuran el sentido del todo actual y porque el todo actual configura el sentido de las partes posteriores, sino porque tanto el todo como las partes son siempre comprendidos por el sujeto poniéndolos en relación con su conocimiento previo sobre el tema del que tratan. De esta manera, cada una de las pistas del juego sólo es informativa para el jugador en tanto que refieren a algo ya conocido por él.

La comprensión como fundamento de la inteligencia y la estructura circular de la comprensión son dos de los temas abordados por el psicólogo e informático Roger Schank para describir cómo funciona la inteligencia a nivel mental. Esta sección vamos a dedicarla a examinar la teoría de Schank. Observaremos que, aunque muestra claras influencias del pragmatismo tal y como revelan sus alusiones a John Dewey (Schank, 1999, p. xi) y su apología del conocimiento procedimental (Ibíd., p 271), la teoría de Schank pertenece no obstante al paradigma cognitivista de la psicología, en tanto que se fundamenta en sus dos supuestos nucleares. Por un lado, el autonomismo reclamado por la tesis internalista, pues ignora la biología del cerebro y del cuerpo. Y por otro, la tesis del procesamiento de información, o metáfora computacional, pues Schank estudia la mente desde un punto de vista molecular, tratando de explicar cómo la inteligencia surge de abajo a arriba, mediante procesos de etiquetado e indización que sirven para almacenar, recuperar y correlacionar jerárquicamente las memorias, a la manera como operan las computadoras.

Scripts

En la década de 1970, mientras trabajaba en el laboratorio de IA de la Universidad de Yale, Roger Schank desarrolló en colaboración con Robert Abelson un programa denominado SAM, acrónimo de *Script Applier Mechanism* (mecanismo aplicador de scripts). SAM era capaz de entender historias escritas, normalmente tomadas de los periódicos, y extractarlas en resúmenes mediante la aplicación de scripts. Un *script*, tal y como lo define Schank, es «una estructura que describe la secuencia apropiada de eventos en un contexto particular o una secuencia predeterminada y estereotipada de acciones que define una situación bien conocida» (Ibíd., p. 8). La idea de fondo de SAM era que para entender una situación o un texto cualquiera no basta con apercibirse de lo que está sucediendo de manera explícita, sino que además es necesario tener un conocimiento sobre lo que sucede de manera implícita y sobre lo que es probable que suceda a continuación. Como ya

mencionamos en el capítulo segundo, el significado de las oraciones emitidas por un sujeto no depende sólo de la semántica de las palabras y del modo en que éstas se combinan de acuerdo a las reglas de la sintaxis, sino que depende también del contexto en el que se profieran, es decir, que depende de la pragmática.

Para ilustrar este fenómeno Marvin Minsky propone la siguiente historia tomada de un texto de su estudiante Eugene Charniak: «Jane fue invitada a la fiesta de cumpleaños de Jack. Ella se preguntó si le gustaría una cometa. Fue a su habitación y agitó su cerdito, pero no sonó» (Minsky, 1975, p. 103). Para comprender el relato, dice Minsky, no basta con conocer el significado de cada uno de sus términos y las reglas de la sintaxis, pues ese conocimiento por sí solo no es suficiente para saber hechos que son imprescindibles para la comprensión. Uno de esos hechos es que cuando uno es invitado a una fiesta de cumpleaños lo normal es llevar un regalo. Sin conocer esta costumbre los eventos "ser invitado a un cumpleaños" y "buscar dinero para hacer un regalo" se presentan como inconexos. Por otra parte, la norma de que los invitados sean los que hacen regalos es válida sólo en nuestro contexto cultural, pues en otras tradiciones lo habitual es lo contrario, y es el que cumple años quien debe regalar algo a sus invitados. Los scripts proporcionaban a SAM este tipo de conocimiento implícito imprescindible para la comprensión. Con los scripts adecuados, SAM habría sido capaz de comprender que el cerdito agitado no era un cerdo de verdad, sino una hucha en forma de cerdo como las que suelen tener los niños para guardar sus ahorros, y la ausencia de sonido indicaba que no había dinero en su interior.

Con el paso de los años Schank se dio cuenta de que el diseño de SAM, aunque parcialmente eficaz, carecía de varias características que son necesarias para conceder que un sujeto comprende lo que lee. En primer lugar, podía suministrársele una y otra vez la misma historia para que la resumiera, y sin embargo no se aburría por la falta de variedad, ni aprendía de las ejecuciones anteriores para efectuar las posteriores con mayor eficacia. Y en segundo lugar, los scripts eran compartimentos estancos que impedían la generalización de conocimientos para aplicarlos en contextos diferentes. Por ejemplo, el script o conocimiento de trasfondo empleado para habérselas con la

situación de comer en un restaurante y el de viajar en avión comparten una escena común que es la de pagar por algo, pero al ser scripts independientes lo que SAM sabía sobre pagar en restaurantes no era capaz de aplicarlo a la situación de pagar billetes de avión. Éstos y otros problemas fueron amontonándose durante años sobre la mesa de Schank, hasta que éste se dio cuenta de que la IA no era un proyecto viable en las próximas décadas, y decidió cambiar de campo de investigación, pasándose de la informática a la psicología del aprendizaje.

Sin embargo, como él señala, en realidad sigue dedicándose a lo mismo de antes, pues «la IA y la educación se preguntan las mismas cuestiones» (Schank, 1999, p. 279). Si no somos capaces de enseñar a los ser humanos para que sean inteligentes, difícilmente podremos enseñar a las máquinas. Para examinar la teoría de la inteligencia de Schank vamos a utilizar su obra *Dynamic memory revisited*, publicada en 1999, la cual consiste en una versión actualizada de las ideas que el autor plasmó en 1982 en *Dynamic memory: a theory of learning in computers and people*, cuando todavía no era consciente de las limitaciones del programa SAM y por ende seguía apostando por la técnica de los scripts estancos como estrategia para diseñar máquinas con la capacidad de comprender textos.

Los psicólogos, dice Schank, suelen ocuparse de la inteligencia sin haber abordado antes el problema de la comprensión debido a que los seres humanos, a diferencia de las máquinas, poseemos esta última facultad de manera natural (Schank, 1986, p. 121). Pero en una investigación rigurosa sobre la inteligencia, la comprensión ha de ser un asunto ineludiblemente previo, dado que la inteligencia es una función del conocimiento (Ibíd., p. 283), y el conocimiento es una parte constituyente de la comprensión. Comprender algo, dice Schank, consiste en relacionarlo con una memoria pertinente: «Encontrar la memoria *adecuada* (la más específica para la experiencia actual) es lo que entendemos por comprender» (Schank, 1999, p. 27). En el caso del relato de Minsky las memorias adecuadas que deben ser elicitadas para operar la comprensión serían las referentes a los usos y costumbres de las fiestas de cumpleaños y la que permite interpretar que el cerdito agitado es una hucha.

Comprender y procesar la información son, tal y como los concibe Schank, términos sinónimos (Ibíd., p. 74). Lo que hace una computadora electrónica cuando procesa la información es relacionarla con otra información pertinente. Por ejemplo, cuando el sistema informático de la administración de hacienda procesa los datos de un contribuyente lo que está haciendo es relacionarlos con los porcentajes de lo que debe pagar por cada uno de los conceptos.

De esta manera, la diferencia de inteligencia entre dos personas estriba en la diferencia de sus capacidades para realizar asociaciones de lo presente con lo ausente rememorado. Decimos que un sujeto es inteligente cuando en una conversación realiza un *salto* imprevisto pero al mismo tiempo atinado. «El salto puede evidenciar una rapidez de pensamiento, comprensión profunda, razonamiento sagaz, o simplemente una perspectiva original sobre un problema cotidiano o una elección de palabras sorprendente. La cuestión es la novedad. Nos impresiona escuchar cosas que son nuevas o diferentes en algún sentido» (Ibíd., p. 233).

Fenomenología de la comprensión

Lo que Schank entiende por el término "comprender" (*understand*) equivale en la teoría fenomenológica de Martin Heidegger a "interpretar". Para el filósofo alemán el *comprender* es un rasgo existencial aún más fundamental que el interpretar, pues refiere a la misma aperturidad del *Dasein*, un término este último que para nuestros propósitos es intercambiable por el de "sujeto cognoscente". «En el por-mor-de está abierto el existente estar-en-el-mundo en cuanto tal; esta aperturidad ha sido llamada comprender» (Heidegger, 1927, p. 167). La *aperturidad* es la característica por la cual el Dasein está abierto a sí mismo y a lo otro. Sin aperturidad no hay interpretación posible porque, sencillamente, no hay apercepción de absolutamente nada, ni de sí mismo ni de lo otro. Las piedras, por ejemplo, no comprenden porque no tienen aperturidad, nada saben de sí mismas ni de lo que las rodea. «El comprender es el ser existencial del propio poder-ser del Dasein mismo, de tal manera que este abre en sí

mismo lo que pasa consigo mismo» (Ibíd., p. 168). Por otra parte, Heidegger señala que la comprensión se produce siempre en un *éxtasis*, o estar fuera de sí, de carácter temporal que unifica los tres horizontes del presente, el pasado y el futuro. Esto quiere decir que el Dasein o sujeto cognoscente se comprende a sí mismo no como un ser aislado en el instante actual indicado por las agujas del reloj, sino como un ser que constantemente presentifica lo que ya no es y lo que todavía no es, es decir, que retiene el pasado y anticipa el futuro.

En cuanto al interpretar, que como decimos equivale a lo que Schank denomina "comprender", es para Heidegger el existencial por el cual el comprender se apropia comprensoramente de lo comprendido (Ibíd., p. 172). *Interpretar* consiste en observar lo abierto a la comprensión "en tanto que" algo. Así, cuando vemos un objeto no vemos un objeto sin más, sino que vemos algo en tanto que una mesa, una puerta o un coche. Lo ahí dado es siempre accesible "en tanto que" debido a que lo abierto por la comprensión se presenta ya interpretado. Para para percibir una puerta como algo que está puramente ahí, es decir, como un objeto no interpretado, es necesario realizar un esfuerzo artificioso de abstracción que, en todo caso, es posterior a la interpretación del objeto como puerta.

El filósofo alemán apunta que el interpretar se funda siempre en un *haber previo*: «La interpretación de algo en cuanto algo ya está esencialmente fundada en el haber previo, en la manera previa de ver y en la manera de entender previa. La interpretación no es jamás una aprehensión, sin supuestos, de algo dado. Cuando esa particular concreción de la interpretación que es la interpretación exacta de los textos apela a lo que "está allí", lo que por lo pronto está allí no es otra cosa que la obvia e indiscutida opinión previa del intérprete, que subyace necesariamente en todo quehacer interpretativo como aquello que con la interpretación misma ya está "puesto", es decir, previamente dado en el haber previo, la manera previa de ver y la manera de entender previa» (Ibíd., p. 174). En el caso de SAM los scripts son el haber previo al que se refiere Heidegger. El interpretar, que es equivalente al comprender según Schank, es por tanto relacionar lo dado con el haber previo, con la manera

previa de ver, con la manera de entender previa. En el interpretar el sujeto cognoscente dota de sentido a lo interpretado, pues sólo el Dasein puede estar dotado de sentido o desprovisto de él. «Sentido es el horizonte del proyecto estructurado por el haber-previo, la manera previa de ver y la manera de entender previa, horizonte desde el cual algo se hace comprensible en cuanto algo» (Ibíd., p. 175).

Ahora bien, dado que el interpretar se basa en lo comprendido previamente, resulta entonces que el comprender se mueve en un círculo en el que nada se presenta tal y como es en sí mismo, sino que el sentido de lo dado procede de lo comprendido con anterioridad. Ver en esta circularidad un círculo vicioso, dice Heidegger, es malcomprender radicalmente el comprender. El comprender es, debido a la estructura temporal extática del Dasein, inevitablemente circular, pues discurre presentificando el pasado y el futuro. La finalidad, por tanto, no es evitar el círculo hermenéutico, sino entrar en él de forma correcta. Esta forma correcta consiste en discriminar y rechazar los prejuicios infundados: «no dejar que el haber previo, la manera previa de ver y la manera de entender previa sean dados por simples ocurrencias y opiniones populares» (Ibíd., p. 176).

Hans-Georg Gadamer, continuador de la tarea de Heidegger en este asunto, apunta que el comprender requiere ineludiblemente la aplicación de *prejuicios*, pero entendiendo éstos en el sentido puramente etimológico de juicios previos, no en el vulgar de juicios populares aceptados de manera acrítica. «Toda interpretación correcta tiene que protegerse contra la arbitrariedad de las ocurrencias y contra la limitación de los hábitos imperceptibles del pensar, y orientar su mirada "a la cosa misma"» (Gadamer, 1960, p. 333). Sin embargo, como ya hemos dicho, la cosa misma no se muestra jamás de manera directa, al modo de la soñada intuición intelectual racionalista, dado que todo comprender es interpretativo, pero puede no obstante ser deducida a partir de la resistencia que muestra a dejarse interpretar por prejuicios erróneos. El que quiere comprender un texto realiza siempre un proyectar de sus prejuicios, dice Gadamer, pero si el texto resulta incomprensible bajo la interpretación resultante de ese proyectar, entonces esos prejuicios rectores se revelan inadecuados

y deben ser sustituidos por otros alternativos. «El que quiere comprender un texto tiene que estar en principio dispuesto a dejarse decir algo por él. Una conciencia formada hermenéuticamente tiene que mostrarse receptiva desde el principio para la alteridad del texto» (Ibíd., p. 335). El buen intérprete ha de suponer, por principio, que el texto al que se enfrenta tiene una *unidad de sentido* y que lo que dice es *verdadero* (Ibíd., p. 363). Estas dos condiciones son las que obligan a buscar un esquema de interpretación tal que la interpretación resultante sea la de un texto con unidad de sentido y verdadero en lo que dice. Se trata de dos principios de pragmática elemental, pues todos suponemos que cuando alguien habla su discurso tiene sentido y pretende decir la verdad. Sólo después de haber fracasado los intentos por dotarle de sentido y de veracidad es justo declararlo incomprensible.

El juego de las 20 preguntas, tan fácil de entender, sirve para aclarar el significado de las difíciles expresiones de Heidegger y Gadamer. En el juego cada nueva pista restringe el número de posibles personajes que encajan con ellas. De manera análoga, un texto, a lo largo de su desarrollo, introduce informaciones que progresivamente van acotando las interpretaciones posibles, mostrándose inaccesible ante aquellas que son contrarias al sentido del discurso que el autor pretende comunicar. De esta manera las partes configuran el sentido de la totalidad. Y a la inversa, el todo proporciona el sentido necesario para interpretar las partes, pues las partes sólo son interpretables sobre el trasfondo de una totalidad, ya sea ésta el haber previo como lo denomina Heidegger o un script como en el caso de SAM. En el juego de las 20 preguntas la hipótesis manejada en cada momento hace las veces de totalidad proporcionadora de sentido para interpretar las partes. En un texto cualquiera sucede lo mismo, pues el sentido global de la obra que hasta un determinado punto de la lectura ha posibilitado la interpretación de lo anterior sirve al lector para generar expectativas de lo que va a suceder a continuación. Si las expectativas resultan fallidas, entonces surge la necesidad de descartar el sentido de la totalidad pensado hasta ese momento y elaborar otro diferente que permita reinterpretar la información recibida como si toda ella tuviera unidad de sentido.

En palabras de Gadamer: «La anticipación de sentido que hace referencia al todo sólo llega a una comprensión explícita a través del hecho de que las partes que se determinan desde el todo determinan a su vez a este todo. Este hecho nos es familiar por el aprendizaje de las lenguas antiguas. Aprendemos que es necesario "construir" una frase antes de intentar comprender el significado lingüística de cada parte de dicha frase. Este proceso de construcción está sin embargo ya dirigido por una expectativa de sentido procedente del contexto de lo que le precedía. Por supuesto que esta expectativa habrá de corregirse si el texto lo exige. [...] El movimiento de la comprensión va constantemente del todo a la parte y de ésta al todo. La tarea es ampliar la unidad del sentido comprendido en círculos concéntricos. El criterio para la corrección de la comprensión es siempre la congruencia de cada detalle con el todo. Cuando no hay tal congruencia, la comprensión ha fracasado» (Ibíd., p. 360).

La conversación (1974) de Francis Ford Coppola es una película que ilustra de forma magistral este proceso circular. En ella un detective privado, interpretado por Gene Hackman, recibe el encargo de un hombre muy poderoso, Robert Duvall, para que espíe los encuentros furtivos de su esposa con el amante de ella. La grabaciones de sonido con micrófonos direccionales que Hackman obtiene de una conversación furtiva entre los amantes son parciales, de tal manera que algunas frases se pierden por culpa del ruido ambiental. Pero a pesar de esas lagunas de información Hackman y el espectador del filme quedan pronto convencidos de que se trata de una pareja de enamorados y que Duvall quiere asesinarlos a ambos, a ella por adúltera y a él por deshonrar su matrimonio. Sin embargo, en los últimos minutos de metraje esta expectativa se trunca, pues se descubre que en realidad eran los amantes los que planeaban matar a Duvall para que su fortuna fuese heredada por ella. Coppola intercala la secuencia final del asesinato de Duvall con unas tomas en las que se escuchan las partes de la conversación de la pareja que no pudieron ser registradas por los micrófonos. En esas alocuciones queda claro que ellos eran los conspiradores. La expectativa ha fracasado, y el espectador ha de elaborar un nuevo sentido global para reinterpretar todo lo que ha sucedido a lo largo de la historia.

En todas las situaciones en las que interviene la comprensión, ya sea la lectura de un texto o pedir comida en un restaurante, hay siempre una información ausente, no captada por los micrófonos por decirlo de manera metafórica, y que el sujeto debe poner de su parte para comprender lo que está sucediendo. Dicha información resulta de la interacción circular entre los prejuicios y el objeto de la comprensión, y a su vez el objeto de la comprensión se muestra tal y como es en tanto que se protege de las interpretaciones incorrectas mediante la exigencia de unidad de sentido en la interacción circular entre el todo y las partes.

Estructura y funcionamiento de la memoria

En definitiva, comprender algo consiste en relacionarlo con memorias pertinentes al respecto. En el caso de la IA fuerte, dichas memorias pueden ser implantadas una a una por los investigadores, o bien la máquina puede ser diseñada de tal forma que sea capaz de elaborar sus propias memorias a partir de la experiencia. El proyecto CYC de Lenat apostó en su día por la primera alternativa, mientras que la segunda es la defendida por Schank y Turing, entre otros muchos expertos en la materia. El *CYC* (del inglés *enCYClopedia*) comenzó su andadura en 1984, en la Microelectronics and Computer Technology Corporation en Texas, con una inversión inicial de 50 millones de dólares (Copeland, 1993, p. 158). A principios de los 90, el CYC se había convertido en el proyecto estrella de la IA simbólica. Su director, Douglas Lenat, pretendía construir una gran base de datos dividida en dos partes (Crevier, 1993, p. 241). Aplicando la terminología de Robert Glaser antes vista, la parte de arriba estaría formada por los *conocimientos artefactuales* recogidos en una enciclopedia, y la de abajo, por un amplio conjunto de enunciados de sentido común que contendrían los *conocimientos naturales* necesarios para entender las entradas de la enciclopedia.

Algunos ejemplos de esos datos de sentido común serían, por ejemplo, que la lluvia cae hacia abajo, que en los restaurantes suele pagarse después de haber comido, y que en las fiestas de cumpleaños en Occidente es cortesía llevar un regalo. El número

de enunciados de este tipo que Lenat estimaba necesario para que el CYC comenzase a ser operativo era de unos cien millones. Tras los seis primeros años de trabajo, la cifra alcanzada era de más de un millón, lo cual representaba, según Lenat, el 0.1% del conocimiento sobre la realidad consensuado en Occidente (Copeland, 1993, p. 160). En la actualidad el proyecto CYC continúa activo, pero tal y como el propio Lenat pronosticó, parece que su finalización llevará muchas más décadas de esfuerzo. La fecha límite para la consumación del proyecto, fijada en un arranque de optimismo para el año 2007, no se ha cumplido (Kaku, 2011, p. 114).

Respecto a la otra opción, la de construir una máquina que elabore sus propias memorias aprendiendo de la experiencia, fue propuesta por Alan Turing en 1950 con estas palabras: «En vez de intentar la producción de un programa que imite la mente adulta, ¿por qué no tratar más bien de producir uno que simule la del niño? Si se le sometiera entonces a una adecuada educación, se obtendría el cerebro adulto. Probablemente el cerebro del niño es como el libro de notas que uno compra en la papelería. Un mecanismo más bien pequeño, y una gran cantidad de hojas en blanco. (Desde nuestro punto de vista, el mecanismo y la escritura son casi sinónimos). Nuestra esperanza consiste en que al ser el mecanismo del cerebro del niño tan pequeño se pueda programar fácilmente algo parecido» (Turing, 1950, p. 43). De esta manera se evitaría la titánica tarea de introducir uno por uno los millones de enunciados de sentido común, pues la máquina los aprendería, al igual que un niño, en tan sólo unos pocos años y sin esfuerzo. En términos de lo que hemos denominado las nuevas gafas de Kant, sólo habría que modelar las lentes innatas, y las adquiridas serían elaboradas por la propia máquina en base a la experiencia.

Schank es un firme defensor de este enfoque de la IA fuerte. Y dado que para enseñar a aprender a una máquina es útil saber primero cómo aprendemos los seres humanos, el psicólogo norteamericano comienza por investigar esto último. Las fases del proceso humano de aprendizaje que él señala pueden ser agrupadas en tres: memorizar, proyectar y evaluar (Schank, 1999, p. 42). Ejecutadas en orden, el proceso de aprendizaje se repite circularmente, una y otra vez (Ibíd., p. 79).

Funcionamiento circular

Podemos introducirnos en el círculo comenzando por la memorización, la cual, en un sentido amplio, consiste no sólo en almacenar los datos entrantes, sino también en etiquetarlos y someterlos a un proceso de producción de generalizaciones. El dato entrante es etiquetado, y gracias a ese índice con sus características esenciales es puesto en relación con otros anteriores semejantes a él, si los hubiera (Ibíd., p. 53). En caso de no haberlos, la entrada es almacenada como la *memoria particular* de un acontecimiento inédito. Si por el contrario los hay, entonces son sometidos a un procedimiento abductivo que extrae de ellos ciertas regularidades, las cuales son almacenadas en forma de *memorias generales* acerca de cómo es el mundo. Por ejemplo, la primera vez que acudimos a un restaurante de comida rápida, dice Schank, la experiencia de tener que pagar antes de comer es un suceso inédito que es almacenado como una memoria particular. Si volvemos a un restaurante de ese tipo, pero no al mismo, entonces esa nueva información es etiquetada y relacionada con la anterior, y ambas sirven como base factual sobre la que se elabora una representación prototípica en forma de memoria general acerca de cómo son los restaurantes de comida rápida. En ocasiones posteriores la memoria general, noción equivalente a la de representación invariable en la terminología de Hawkins, sobre los restaurantes de comida rápida será utilizada cada vez que entremos en uno de esos establecimientos para proyectar generalizaciones sobre lo que no vemos pero está sucediendo en ese contexto y para predecir lo que sucederá a continuación.

La proyección en forma de generalizaciones y de predicciones es la utilidad fundamental de la memoria, dice Schank (Ibíd., p. 55), respaldando así la tesis de Hawkins (Hawkins & Blakeslee, 2004, p. 66). Por último, si en la fase de evaluación se determina que la proyección ha resultado exitosa, entonces ésta es reforzada inductivamente como una norma válida, y se consolida su utilización en el futuro (Schank, 1999, p. 46). De lo contrario, se produce un fallo que pone en marcha un

proceso explicativo (Ibíd., p. 152). La conclusión de éste puede ser que se cometió un error en la selección de la representación invariable, en tanto que la escogida no era la pertinente, o bien que la representación invariable sí era la pertinente y en consecuencia debe ser modificada para dar cuenta de la desviación acontecida respecto de la norma. En tal caso puede ser necesario acumular más desviaciones similares para elaborar abductivamente una nueva norma que será almacenada en forma de memoria general.

En este punto debemos señalar dos cuestiones. La primera es que lo que Schank denomina como *memorias particulares* y *memorias generales* son conceptos que parecen identificarse respectivamente con los ya descritos de *memorias episódicas* y *memorias semánticas*. Sin embargo, Schank rechaza esta correlación de pares por considerar que la distinción tradicional entre memoria episódica y semántica es problemática en tanto que ambos conceptos no están bien definidos (Ibíd., p. 102). En su lugar maneja las nociones de memorias particulares, o *estructuras memorísticas basadas en historias* (*story-based*), y memorias generales, o *estructuras memorísticas basadas en eventos* (*event-based*).

La unidad del conocimiento

La segunda cuestión que debemos examinar consiste también en una coincidencia, en este caso sin objeciones, entre las fases del proceso de aprendizaje según Schank y las tres estrategias lógicas empleadas en la producción de conocimiento científico que vimos en el capítulo anterior: abducción, deducción e inducción. Así, en el contexto de descubrimiento la memorización contiene un momento de *abducción* que transforma las memorias particulares en generales, mientras que en el contexto de justificación la deducción proyecta consecuencia particulares a partir de memorias generales, y la inducción, a pesar de su falta de validez lógica, consolida las memorias generales exitosas. Abducción, deducción e inducción son justamente las tres formas de inferencia distinguidas por Charles S.

Peirce: «La deducción prueba que algo debe comportarse de una forma determinada; la inducción que algo se comporta fácticamente así, y la abducción que presumiblemente algo se comporta así» (Habermas, 1968a, p. 120).

La coincidencia entre los tres *subprocesos del aprendizaje* distinguidos por Schank y las tres *estrategias lógicas* empleadas en la producción de conocimiento científico se debe a que, como dice Jürgen Habermas, las ciencias naturales, denominadas por él como ciencias empírico-analíticas, no son más que «la continuación sistemática de un proceso de aprendizaje acumulativo que se realiza de forma precientífica en el ámbito funcional de la actividad instrumental» (Ibíd., p. 194). En cuanto a las *estrategias psicológicas* del contexto de descubrimiento científico, de las cuales reseñamos anteriormente la serendipia y el razonamiento por analogía, también están presentes en el aprendizaje. En lo referente a la serendipia, se trata de una manera de descubrir explicaciones que acontece cuando experimentamos por ensayo y error, en ausencia habitualmente de una expectativa de lo que va a suceder a continuación. Y acerca del razonamiento por analogía, Schank observa que «la mayoría de la gente realiza las explicaciones copiando otras que han sido utilizadas en otros lugares. Ciertamente, podrían considerar muchas posibilidades, ponderar las diferencias entre ellas y entonces alcanzar conclusiones más válidas. Pero, de hecho, casi nunca lo hacen. [...] Permiten a sus prejuicios acerca de por qué la gente hace las cosas restringir la búsqueda de una explicación original. Aunque esto puede parecer una suerte de pereza cognitiva, en realidad no hay mucho tiempo en el curso del procesamiento de una situación para contemplar todas las explicaciones posibles» (Schank, 1999, p. 67). Los marcadores somáticos, como dice Damasio, restringen el elenco de explicaciones evaluables de manera consciente para permitir así al sujeto la toma rápida de decisiones, pues la velocidad de respuesta supone a menudo para los animales la diferencia entre la vida y la muerte.

Así pues, en virtud de la continuidad epistemológica entre el proceso natural de aprendizaje y el de elaboración de teorías científicas que acabamos de señalar de la mano de Habermas, y dada la unidad del conocimiento que hemos proclamado en el

capítulo anterior al denunciar el mito del método, se colige la tesis de que el problema de formalizar algorítmicamente la producción de conocimiento científico tanto natural como social es, en el fondo, el mismo problema que el de crear inteligencias artificiales simbólicas en sentido fuerte. Por tanto, en el momento en que el proceso de generación de enunciados científicos fuera formalizable algorítmicamente, entonces la IA fuerte sería una realidad. Y a la inversa, si se lograse construir una auténtica IA fuerte, entonces bastaría con proporcionarle cantidades ingentes de experiencia para que ella postulara y corroborase millones de posibles teorías científicas. En el caso de que la experiencia a su alcance no fuese suficiente para falsar una hipótesis o para sugerirle determinados indicios inductivos, entonces la máquina pediría a los seres humanos que ellos se encargasen de realizar los experimentos necesarios. Puede parecer una idea bizarra, o una veleidad de la ciencia ficción, pero la IA fuerte generadora de teorías científicas sería la perfecta culminación de la máxima acuñada por Gaspard De Prony de «fabricar logaritmos igual que se fabrican agujas» (Guijarro & González, 2010, p. 182), es decir, de la voluntad de producir conocimiento de manera automatizada, como también pretendía David Hilbert en las matemáticas. Nos ocuparemos de este asunto en el capítulo octavo, cuando examinemos las condiciones de posibilidad sociales de la IA fuerte.

Estructuras memorísticas

Volviendo al texto de Schank, encontramos que distingue un tipo de memoria particular, que son los scriptlets, y tres tipos de memorias generales: escenas, MOPs y TOPs. Estas cuatro *estructuras memorísticas* (*memory structures*) se ordenan así de menor a mayor grado de abstracción, o al revés, de mayor a menor grado de concreción: scriptlets, escenas, MOPs y TOPs. Cada una de ellas no sólo es una estructura de almacenamiento de información, sino también de procesamiento de la misma (Schank, 1999, p. 27), de manera que la distinción de la arquitectura von Neumann entre ambos tipos de estructuras no es aplicable, y la distinción equivalente

entre conocimientos y procesos trazada por Sternberg a lo largo de su esquema sobre los lugares de la inteligencia antes expuesto sólo resulta válida a efectos analíticos (Sternberg, 1986, p. 4). Schank define un *scriptlet* como el recuerdo de un episodio particular (Schank, 1999, p. 118). Por ejemplo, aquella ocasión en la que entramos en un restaurante de comida rápida por vez primera y, sorprendentemente, nos hicieron pagar antes de comer. Dado que la memoria es un sistema que tiende a almacenar generalizaciones y a desechar las particularidades (Ibíd., p. 174), los scriptlets que se recuerdan suelen ser aquellos que se refieren a situaciones de gran relevancia emocional o en las que una expectativa importante y muy asentada no se cumplió (Ibíd., p. 101). De esta manera, es probable que se recuerde la primera vez que sucedió la situación imprevista del restaurante de comida rápida, pero las que vinieron después son olvidadas una vez han cumplido con su función de servir para abducir la norma general acerca de lo que suele acontecer en los restaurantes de ese tipo.

En cuanto a las memorias generales, una *escena* (*scene*) es «una estructura memorística que agrupa un conjunto de acciones las cuales comparten una meta y suceden al mismo tiempo» (Ibíd., p. 125). Las escenas se generan abductivamente a partir de la repetición de scriptlets similares (Ibíd., p. 157). Por su parte, un *MOP*, acrónimo en inglés de *paquete de organización de memoria* (*memory organization packet*) consiste en «un conjunto de escenas dirigidas hacia la consecución de una meta. Un MOP siempre tiene una escena principal cuya meta es la esencia o el propósito de los eventos organizados por el MOP» (Ibíd., p. 123). Por tanto, la diferencia entre una escena y un MOP es que este último es una estructura memorística más abstracta que contiene etiquetas que señalan hacia escenas. Schank reconoce moverse en un terreno difuso, en tanto que «no hay una respuesta correcta sobre qué puede ser una escena o un MOP. Las entidades empleadas por una memoria varían en función de las entradas que han sido procesadas y de las generalizaciones que han sido hechas» (Ibíd., p. 131). Así, lo que para una persona puede ser una escena, para otra puede ser un MOP descomponible en escenas más simples. Depende de la agudeza mental de cada uno.

Por ejemplo, se puede convenir que habitualmente el MOP M-RESTAURANTE contiene índices que apuntan de manera ordenada a las escenas PEDIR, COMER y PAGAR. De esta manera, la escena PAGAR puede ser invocada por otros MOPs diferentes, a diferencia de lo que sucedía en la antigua teoría de los scripts de Schank, en la que éstos eran estructuras cerradas que no permitían el intercambio de información. Así, lo que aprendemos acerca de PAGAR en los restaurantes puede ser empleado por MOPs como M-VISITA AL DENTISTA o M-VIAJAR EN AVIÓN. A su vez, la escena PAGAR puede ser instanciada por el scriptlet $*tarjeta de crédito*, con sus respectivas desviaciones de la norma que pueden suceder, tales como "límite de crédito excedido" o "sólo aceptan dinero en metálico" (Ibíd., p. 125). En el caso de la primera visita a un restaurante de comida rápida, el imprevisto de tener que pagar antes de comer supuso un incumplimiento de la secuencia de eventos contenida en M-RESTAURANTE. Esta desviación de la norma, sumada a las posteriores visitas a otros restaurantes de ese tipo, dio lugar a la creación de un nuevo MOP, M-FASTFOOD, en el cual el orden de las escenas es PEDIR, PAGAR, COMER.

Finalmente, la estructura memorística más abstracta y de más alto nivel de todas en la teoría de Schank es el *TOP*, siglas de *paquete de organización temática* (*thematic organization packet*). Mientras que los MOPs son relativos a un dominio específico, los TOPs codifican conocimiento intercontextual, aplicable por tanto a cualquier dominio (Ibíd., p. 145). Un ejemplo de TOP sería el concepto "imperialismo", aplicable a dominios tan dispares como la política, la economía o incluso a una relación interpersonal en la que alguien exhibe un carácter dominante. Los proverbios, dice Schank, son un tipo muy claro de TOP, pues se trata de sentencias proporcionadas por el acervo de la cultura popular que contienen conocimientos explicativos de una generalidad máxima: más vale pájaro en mano que ciento volando; antes se atrapa al mentiroso que al cojo; una golondrina no hace primavera. Al igual que en el caso de los scriptlets, las escenas y los MOPs, también los TOPs se caracterizan porque pueden presentarse varios de ellos simultáneamente. Así, el concepto de imperialismo puede venir acompañado del proverbio "torres más altas han caído".

Uno de los aspectos más interesantes del esquema organizativo de la memoria propuesto por Schank es que las estructuras memorísticas suelen presentarse en grupos de tres. Aunque también puede suceder en los scriptlets, las escenas y los TOPs (Ibíd., p. 155), son sobre todo los MOPs los que se presentan en tríos, conteniendo uno de ellos la *información física*, otro la *información social* y un último la *información personal* o *idiosincrásica* (Ibíd., p. 115). De esta manera, la visita al médico «puede ser entendida, almacenada y rememorada en términos de aspectos físicos –conducir, esperar, ser examinado, marcharse, y demás. También puede ser entendida en términos de aspectos sociales, en este caso el acuerdo tácito de pagar por los servicios recibidos. Y adicionalmente, la visita puede ser entendida en términos de las diversas metas personales que están presentes por parte de los participantes del evento. En el caso del paciente, sería M-PRESERVACIÓN DE LA SALUD. En el del médico, M-TRABAJO controlaría sus acciones» (Ibíd., p. 124). Esta pluralidad de dimensiones refleja la división de la inteligencia específica en los dominios físico y social (García, 2001a, p. 176), tal y como venimos haciendo desde el primer capítulo, pudiendo considerarse que la tercera en liza, la dimensión de la información personal, está en el fondo ligada a todo tipo de memorias, ya sean físicas o sociales (Schank, 1999, p. 124).

Niveles de comprensión

Siendo la teoría de la comprensión de Roger Schank tal y como la hemos descrito a grandes rasgos, el psicólogo norteamericano distingue tres niveles de comprensión, que de menor a mayor son: tener sentido (*making sense*), comprensión cognitiva (*cognitive understanding*) y empatía completa (*complete empathy*) (Schank, 1986, p. 124). Un suceso *tiene sentido* «cuando los eventos que ocurren en el mundo pueden ser interpretados por el sujeto en términos de una representación coherente (aunque probablemente incompleta) de cómo sucedieron esos eventos en el pasado» (Ibíd., p. 122). Por ejemplo, podemos decir que un texto tiene sentido para alguien si es capaz de resumirlo o de traducirlo a otro idioma. En el otro extremo, la *empatía*

completa se produce cuando un sujeto conoce de antemano o es capaz de inferir casi a la perfección las representaciones mentales y los estados emocionales que subyacen a la conducta de otra persona. Y, entre medias, la *comprensión cognitiva* tiene lugar cuando no hay una empatía completa pero no obstante hay un conocimiento más profundo que el del mero tener sentido. Siguiendo con el ejemplo del texto, habría comprensión cognitiva si el lector fuera capaz de formular hipótesis acertadas sobre las causas de los sucesos descritos (Ibíd., p. 124).

El grado de comprensión es, en opinión de Schank, un indicador del grado de inteligencia (Ibíd., p. 127). Mientras que comprender algo es darle sentido relacionándolo con memorias pertinentes, la *inteligencia* es la habilidad de retener la experiencia, de abstraer de ella reglas explicativas generales y de utilizar éstas para la comprensión: «De hecho, lo que consideramos como diferencias individuales en inteligencia pueden no ser más que diferencias en la capacidad para hacer abstracciones que crean organizaciones de memoria que permiten observaciones perspicaces no disponibles para aquellos que no han logrado hacer dichas abstracciones» (Schank, 1999, p. 142).

El problema del conocimiento

En el presente análisis de la teoría de Schank, la última tarea que nos resta por abordar es la de evaluar su utilidad de cara a la construcción de inteligencias artificiales fuertes de tipo simbólico. Una manera de hacerlo es comprobando hasta qué punto resuelve el denominado *problema del conocimiento*. Con este nombre se conoce al triple problema de organizar, actualizar y extraer los datos de la memoria de una computadora electrónica (Copeland, 1993, p. 142). Por lo menos a modo de estrategia argumentativa, concedamos a Schank que sus tesis sobre la organización y la actualización son válidas, y que las describe con el detalle suficiente para formalizarlas en algoritmos ejecutables por una computadora electrónica. Así, en cuanto a la *organización*, la jerarquización vertical de las estructuras memorísticas de menor a

mayor grado de abstracción y la transversalidad horizontal entre ellas articulada mediante la intercambiabilidad de las escenas entre MOPs y la intercontextualidad de los TOPs son características que parecen reflejar la organización de la memoria humana. Respecto a la *actualización*, o *problema del marco* (*frame problem*), las tesis de Schank sobre el modo en que el cumplimiento y las desviaciones de la norma afectan a las estructuras memorísticas también parecen dar buena cuenta de cómo se generan y modifican las memorias en la mente humana. Pero en lo que se refiere al tercer y último problema, el de la *extracción* de los datos, más conocido como el *problema de la selección del marco* (*frame selection problem*) o *problema de la pertinencia*, Schank no es capaz de aportar soluciones concluyentes (Dreyfus, 1992, p. 45). La pregunta es: ¿Cuál es el proceso por el cual en la situación del cumpleaños antes mencionada es elicitado el MOP M-CUMPLEAÑOS con sus escenas y scriptlets asociados, y no cualquier otro como M-RESTAURANTE?

De nuevo, nos encontramos ante el *problema de los universales*: cómo es posible reconocer que un objeto particular pertenece a una categoría universal respecto de la cual no es idéntico. En palabras de Hubert Dreyfus, uno de los autores más críticos con la IA simbólica: «Para reconocer un contexto uno debe haber seleccionado previamente las características relevantes de entre el número indefinido de características, pero esa selección sólo puede ser realizada una vez el contexto ha sido reconocido como similar a otro ya analizado» (Dreyfus, 1979, p. 200). A nivel cerebral, o subsimbólico, la solución, tal y como apuntaba Jeff Hawkins, reside en la combinación del principio de Hebb con la estructuración de las redes neuronales en una jerarquía de abstracción creciente y con las conexiones de realimentación. Sin embargo, a nivel mental, o simbólico, todas las tentativas de los investigadores de la IA, desde Minsky hasta Schank, han fracasado.

La dificultad radica en descomponer en una secuencia de eventos sucesivos lo que en realidad son dos procesos opuestos que acontecen sin que ninguno de ellos pueda ser anterior al otro. Esto es: cada uno es la condición de posibilidad del otro. Así como en un movimiento circular físico actúan dos fuerzas contrarias y simultáneas, la

centrífuga y la centrípeta, en el movimiento circular de la comprensión las dos fuerzas opuestas presentes son los dos grandes tipos de pensamiento distinguidos por la tradición psicológica: el análisis y la síntesis (García, 1996, p. 303). Ambos aparecen ya recogidos en Hobbes, quien dice de la justicia o *análisis* que es la facultad de apreciar diferencias, mientras que el ingenio o la *síntesis* es la de apreciar semejanzas (Hobbes, *Leviatán*, p. 68). El movimiento circular de la comprensión implica un análisis para distinguir el objeto mediante su clasificación en una categoría, pero a su vez la distinción sólo es posible sobre la síntesis o semejanza del objeto con una categoría.

De entre las muchas diferencias que pueden observarse entre una computadora electrónica corriente y una IA fuerte, dice Schank, una de las principales es que la IA fuerte debe comprender la realidad de manera *activa*, proyectando expectativas tal y como hacemos los seres humanos (Dreyfus, 1979, p. 180). Por tanto el problema de la selección de los marcos, que son las estructuras generadoras de expectativas, es prioritario. En línea con la continuidad que venimos subrayando entre el conocimiento científico y el precientífico, Marvin Minsky define la función de los marcos en la IA como análoga a la de los paradigmas en la teoría de Thomas Kuhn (Minsky, 1975, p. 97). En palabras del propio Kuhn: «En ausencia de un paradigma o de algún candidato a paradigma, todos los hechos que podrían estar relacionados con el desarrollo de una ciencia posiblemente se presentarán como igualmente relevantes» (Kuhn, 1970, p. 15). Es el marco o paradigma el que discrimina lo relevante de lo irrelevante, lo pertinente de lo no pertinente. Comprender no consiste simplemente en relacionar el objeto de la comprensión con memorias cualesquiera, sino con memorias *pertinentes*, pero a su vez esa pertinencia sólo es determinable una vez que las características esenciales del objeto de la comprensión han sido destacadas por encima de las demás, lo cual sólo puede suceder como resultado de un acto comprensivo previo.

Supongamos que un sujeto escucha una historia. En la siguiente cita Schank condensa su intento por solucionar el problema de la selección del marco: «La relevancia es establecida correlacionando las características de la historia escuchada

con características de historias que ya están presentes en la mente del sujeto. Para realizar esto, el sujeto debe descomponer lo que escucha en unos términos que sean precisamente los mismos que utilizó anteriormente para almacenar situaciones sobre las que había oído hablar o que él mismo había experimentado. En cualquier caso, la correlación es efectuada sobre la base de que estos términos comunes sirven como índices en la mente» (Schank, 1999, p. 226). Si la historia fuese, por ejemplo, la de Romeo y Julieta, Schank distingue en ella los siguientes *índices*: en cuanto a la *meta*, el índice sería la persecución de los personajes de un objetivo común; en cuanto a las *condiciones*, la oposición del entorno; y en cuanto a las *características*, entendiendo por tales los rasgos más contingentes, los índices serían una pareja de amantes jóvenes y el descubrimiento de una muerte falsa (Ibíd., p. 140).

Ciertamente, empleando índices es fácil explicar cómo la historia de Romeo y Julieta acude a la mente de un espectador a modo de marco para permitirle comprender la película *West side story* (1961) que está viendo en ese momento. Ahora bien, esta estrategia de Schank sólo desplaza el problema de la selección del marco, pues a continuación debería explicar cómo funciona el proceso de generación de índices, que no es más que el proceso de abstracción de las características esenciales de un objeto, es decir, de aquellos rasgos que lo hacen ser lo que es. En este punto, el psicólogo norteamericano no dispone de una respuesta concluyente, debido a que la creación de índices es un proceso inconsciente, y en consecuencia de difícil acceso.

Procesos inconscientes

De hecho, tanto la organización, como la actualización y la extracción del conocimiento, son procesos que se realizan casi todo el tiempo de manera *inconsciente*, sin que el sujeto se aperciba de ellos. De lo único de lo que somos conscientes es de lo que sucede en nuestra memoria de trabajo (Carr, 2010, p. 123), y la memoria de trabajo tiene una capacidad muy reducida. Por tanto, la mayor parte de los procesos mentales suceden más allá de la supervisión de nuestras facultades

metacognitivas. Prueba de ello, dice Schank, es que «no podemos verbalizar qué es lo que estamos haciendo cuando nos cepillamos los dientes, conducimos el coche, tiramos o cogemos una pelota, masticamos y tragamos, construimos una oración, o entendemos la televisión» (Schank, 1999, p. 196). Sin embargo, desde Platón la sabiduría en Occidente se ha identificado con la verbalización. Para el filósofo griego la educación (παιδεια) consiste en traer a la conciencia mediante el verbo (λόγος) el conocimiento inconsciente contenido en el alma, el cual habría sido adquirido previamente por la contemplación de las Ideas durante el viaje celeste. La apología racionalista del verbo sigue vigente en nuestra época, denuncia Schank, y es, junto con el legado de la pedagogía conductista (Ibíd., p. 270), la causa del fracaso escolar y de la falta de competencias reales de quienes se gradúan. El sistema educativo está diseñado para que los alumnos adquieran *memorias declarativas* que les permitan verbalizar de manera consciente *qué* son las cosas, una pretensión que es contraria a nuestras disposiciones naturales, pues el cerebro ha sido modelado durante milenios de selección natural para servir a la supervivencia, y para sobrevivir el conocimiento necesario es el práctico de las *memorias procedimentales* que contienen información acerca del *cómo*, ya sea cómo transformar el medio, cómo adaptarse a él, o cómo buscar otro diferente; los tres objetivos de la inteligencia según la subteoría contextual de la teoría triárquica de Sternberg.

El verdadero conocimiento «no es la ilustración cultural o la habilidad de exhibir lo que sabes, sino la habilidad de demostrar lo que puedes hacer. No hay mucha diferencia entre la inteligencia humana y la de los animales superiores. Darwin lo señaló de varias maneras. Un gato, por ejemplo, puede fracasar en sus expectativas, y entonces reacciona elaborando generalizaciones que le permitan cambiar su conducta en respuesta a ese tipo de eventos. Puede rememorar situaciones, o al menos parece que recuerda situaciones cuando se encuentra con otras similares» (Ibíd., p. 271). La sabiduría del *qué* es exclusiva de la consciencia, mientras que la del *cómo*, aun cuando haya sido adquirida inicialmente a través de la consciencia, sólo alcanza su grado óptimo de eficacia cuando se ejecuta de manera inconsciente. Véase

la diferencia de destreza al volante entre el veterano que conduce con tanta facilidad como camina y el novato que necesita pensar en todos sus actos y por tanto comete numerosos errores.

En relación a la IA, Schank apunta que «conseguir que una máquina adquiera conocimiento es el asunto clave. Sin embargo, el conocimiento factual es de poca utilidad para una máquina. El conocimiento procedimental, de saber *cómo hacerlo* en vez de saber *qué hacer*, es la diferencia entre las máquinas inteligentes y las no inteligentes» (Ibíd., p. 209). El problema es que, dada la naturaleza inconsciente de la mayor parte de nuestros procesos cognitivos, ni siquiera nosotros mismos sabemos cómo hacemos lo que hacemos. «Lo que yo creo acerca de lo que yo sé, y lo que de hecho yo sé, no son la misma cosa» (Ibíd., 230). Para programar un sistema experto de, por ejemplo, diagnóstico médico, los ingenieros informáticos interrogan a varios médicos con la esperanza de que esas entrevistas revelen cómo se diagnostica a un paciente. Pero la realidad es que los médicos no son capaces de verbalizar cómo se realiza un diagnóstico. Saben hacerlo, pero no expresarlo conscientemente. De manera similar, todos sabemos subir escaleras, pero verbalizar esa memoria procedimental requiere de un esfuerzo ímprobo y artificioso que, además, se revela absurdo en tanto que es insuficiente para enseñar a alguien a subir escaleras, tal y como reflejó Julio Cortázar en su célebre relato *Instrucciones para subir una escalera*.

Las memorias explícitas o declarativas que los médicos adquieren en la universidad mediante el estudio de los libros no es suficiente para realizar un diagnóstico, razón por la cual necesitan superar varios años de prácticas para obtener el graduado. Y a la inversa, las memorias implícitas o procedimentales que adquieren durante esos años de prácticas no son capaces de expresarlas declarativamente para que el ingeniero informático las codifique en algoritmos y se las proporcione a la computadora electrónica. El resultado es que los sistemas expertos cometen errores, algunos de los cuales van más allá de la falta de conocimiento específico sobre la medicina o el área de que se trate, y revelan una grave carencia de conocimiento de sentido común. En palabras de Douglas Lenat: «Puesto que carecen [...] de conceptos

simples del sentido común, los errores de los sistemas expertos parecen ridículos en términos humanos. Por ejemplo, cuando un programa de financiación de coches aprueba un préstamo para un adolescente que escribe que ha estado en el mismo trabajo durante veinte años; [...] o cuando un sistema médico prescribe una dosis absurda de una droga para una paciente de la Maternidad cuyo peso (80) y edad (35) fueron intercambiados accidentalmente al tomar los datos» (Copeland, 1993, p. 159).

En resumen, aun suponiendo a modo de estrategia argumentativa que las tesis cognitivistas de Schank acerca de la forma en que el conocimiento se organiza y actualiza en la memoria fuesen válidas y pudieran ser formalizadas algorítmicamente, para resolver el problema del conocimiento, crucial para la IA simbólica, necesitaría dar cuenta también del proceso de extracción. La coincidencia de índices, que es la esencia de su propuesta, supone la capacidad para indizar, es decir, formar etiquetas contenedoras de los rasgos esenciales, pero los rasgos esenciales sólo son destacables respecto de los demás en presencia de un marco, y la selección del marco requiere de una previa coincidencia de índices. Por tanto, la estructura circular de la comprensión, que parece replicable a nivel cerebral mediante circuitos de realimentación tal y como hemos visto en la teoría de Hawkins, en el mental se resiste a un enfoque cognitivista como el de Schank. Volveremos a este asunto en el capítulo séptimo.

5.2.3. La inteligencia en la conducta

En esta tercera sección, dedicada a la inteligencia en la conducta, vamos a examinar la *teoría de las inteligencias múltiples*, o *teoría IM*, de Howard Gardner, quien recordemos que define la inteligencia como «la capacidad para resolver problemas, o para elaborar productos que son de gran valor para un determinado contexto comunitario o cultural» (Gardner, 1993, p. 27). El psicólogo norteamericano denuncia con pesadumbre que sus colegas de profesión, lejos de arriesgarse a proponer definiciones informativas como la suya, suelen referirse a la inteligencia de manera operacional como «la habilidad para responder a las cuestiones de un test de

inteligencia» (Ibíd., p. 37). Los tests, dice Gardner, han sido tradicionalmente construidos sobre una noción unitaria, descontextualizada e individual de la inteligencia, mientras que la de su teoría él la caracteriza como plural, contextualizada y distribuida. Veamos lo que significan estas tres propiedades.

Plural

Que la inteligencia es *plural* refiere a que no hay una sola inteligencia general encargada de resolver todos los problemas o de elaborar todos los productos de valor cultural, sino que hay varias inteligencias, las cuales en opinión de Gardner funcionan sin un *yo* ejecutivo central (Ibíd., p. 70) y suelen operar de forma conjunta incluso para acometer las tareas más simples (Ibíd., p. 39). Para que una habilidad sea considerada como una inteligencia ha de satisfacer al menos la mayoría de los siguientes criterios (Gardner, 1999, p. 16): poder ser aislada por lesiones cerebrales, existir individuos que la posean en grado excepcional, tener un conjunto de operaciones nucleares, poder seguir su desarrollo y especialización a lo largo de la vida de un sujeto, haber registros de su Historia evolutiva en las especies, disponer del apoyo experimental de pruebas psicológicas, que las baterías psicométricas arrojen resultados concluyentes sobre su identidad y, por último, ser codificable en un sistema de símbolos.

Por poner algunos ejemplos en el mismo orden: las lesiones en el área de Wernicke, situada en el área 22 de Brodmann, dan lugar a la pérdida o disminución de la facultad para entender el lenguaje sin que ello afecte a otras competencias; los deficientes geniales con síndrome de *savant* tales como Kim Peek, en cuya vida se basó la película *Rainman* (1988), tienen algunas habilidades en un grado extraordinario al tiempo que son incapaces de realizar las tareas cotidianas más sencillas; la habilidad para entonar bien es una de las operaciones nucleares de la competencia musical; en la trayectoria de un atleta se puede observar cómo van mejorando sus destrezas cinético-corporales; la inteligencia espacial está presente en todos los mamíferos; hay diversas tareas que demuestran en condiciones de laboratorio qué habilidades están

relacionadas entre sí y cuáles son independientes; las baterías de pruebas psicométricas revelan qué tareas reflejan el mismo factor subyacente; y, por último, los pentagramas son uno de los sistemas simbólicos en los que se codifica la competencia musical. Utilizando estos criterios Gardner ha distinguido a lo largo de su obra un número variable de inteligencias, pero suele considerarse que las principales han sido siempre estas siete: musical, cinético-corporal, lógico-matemática, lingüística, espacial, interpersonal e intrapersonal. Cada una es un potencial psicobiológico capaz de resolver problemas o de elaborar productos valiosos para una comunidad.

Entre algunos profesionales de la psicometría, dice Gardner, perdura sin embargo la antigua creencia de que existe un solo potencial psicobiológico, el cual intentan medir con sus tests (Gardner, 1993, p. 220). Es el caso de Hans Eysenck, quien, como ya hemos mencionado, sólo considera posible abordar desde un punto de vista científico la inteligencia fluida (G*f*) o inteligencia A de Hebb, que es la que depende de las propiedades fisiológicas del cerebro, al tiempo que descarta la inteligencia cristalizada (G*c*) o inteligencia B por ser el producto de la interacción de una cantidad demasiado numerosa de factores ambientales complejos (Eysenck, 1986, p. 69). Eysenck pertenece a una tradición de la psicometría más cercana al planteamiento fisicalista de Galton que al mentalista de Binet. En cuanto a los profesionales de la psicometría que adoptan una perspectiva plural, Gardner estima que sus teorías no reflejan la verdadera composición de la inteligencia debido a que se basan en el análisis factorial. John Carroll describe el *análisis factorial* como «una técnica para descubrir las dimensiones de las habilidades reveladas en los tests psicológicos y la estructura de esas dimensiones» (Carroll, 1986, p. 52). La gran deficiencia del análisis factorial, dice Gardner, es que «los resultados obtenidos son un reflejo directo de las hipótesis matemáticas asumidas al definir y aislar (más técnicamente, "alternar") los factores» (Gardner, 1993, p. 285). De manera que en función de las hipótesis iniciales el resultado puede ser tanto una visión unitaria como plural de la inteligencia, pudiendo obtener desde los escasos 7 factores de la teoría de Louis Thurstone hasta los 120 de la de Joy Guilford.

Contextualizada

El segundo atributo característico de la inteligencia según Gardner es que es *contextualizada*. «Desde el momento en que se valora una capacidad en una cultura, puede considerarse una inteligencia; pero en ausencia de una aprobación de estas características, una capacidad no puede considerarse inteligencia» (Ibíd., p. 84). La inteligencia, por tanto, no debe ser identificada con la sola competencia del individuo, sino que dicha competencia debe ser contemplada en conexión con los dominios y los ámbitos en los que se presenta en la realidad. Gardner utiliza en este punto la distinción entre individuo, dominio y ámbito, inspirada en el trabajo de sus colegas David Feldman y Mihály Csikszentmihalyi (Ibíd., p. 314). En esta terna de conceptos el individuo es el poseedor de la *inteligencia* (*intelligence*) en el sentido de potencial psicobiológico, la *especialidad* o *dominio* (*domain*) es la disciplina o el arte que se practica en una sociedad determinada, y el *ámbito* (*field*) es el conjunto de instituciones y jueces que determinan qué productos son válidos dentro de una sociedad, cómo se construyen los dominios y cuáles son las inteligencias más apreciadas dentro de cada dominio (Ibíd., p. 64).

Así pues, el dominio y el ámbito son los sistemas dinámicos que contextualizan la inteligencia. Para hablar de inteligencia, tan necesario como el potencial psicobiológico del individuo es que dicho potencial tenga algún dominio de aplicación y que esa aplicación sea valorada por instituciones sociales. Por el contrario, considerar que el potencial psicobiológico es por sí solo constitutivo de la inteligencia o que el conjunto de dominios y ámbitos de Occidente en la actualidad son universales son dos errores que, en opinión de Gardner, abundan en los tests. De nuevo Eysenck sería un ejemplo claro para Gardner del investigador que identifica la inteligencia con el potencial psicobiológico por sí solo. En cuanto a la práctica etnocéntrica de considerar que los dominios y ámbitos actuales de Occidente son universales, da lugar a paradojas como la revelada por el efecto Flynn.

Basándose en estudios surgidos en los 80, el psicólogo norteamericano James Flynn observó en 2007 que las puntuaciones de cociente intelectual o CI obtenidas por los escolares de su país en los tests de inteligencia habían aumentado de manera sostenida a lo largo del siglo XX. Este fenómeno fue bautizado como el *efecto Flynn*, y para explicarlo surgieron numerosas teorías (Carr, 2010, p. 146). Algunas afirmaban que se debía a la mejora en la alimentación, otras que la causa residía en una disminución de la natalidad que permitía a las familias concentrar sus recursos en un número menor de hijos, y algunas más creían que la explicación se encontraba en el aumento progresivo de las tasas de escolarización. Por su parte, Flynn advirtió que si esa mejora de los resultados era real, entonces habría que concluir forzosamente que nuestros antepasados de principios de siglo eran retrasados mentales: «La paradoja del retraso mental: si proyectamos los aumentos de CI a 1900, el CI medio de aquella época calculado con las normas actuales estaba entre 50 y 70. Si los aumentos de CI son reales en algún sentido, estamos abocados a la conclusión absurda de que la mayoría de nuestros ancestros eran retrasados mentales» (Flynn, 2007, p. 9). Flynn argumentó que lo que en realidad había sucedido durante el último siglo es que había acontecido un cambio en la convención social de lo que se consideraba ser inteligente, es decir, que las instituciones sancionadoras de los ámbitos habían cambiado su parecer acerca de cuáles eran los dominios más importantes de la vida y cuáles eran las inteligencias que intervenían en ellos.

A principios del siglo XX los padres de la psicometría, con Binet a la cabeza, sentaron las bases de los tests de inteligencia como exámenes de competencias intelectuales de tipo académico, considerando que las habilidades lógico-matemáticas y lingüísticas eran las más importantes. Sin embargo, en aquella época las competencias necesarias para la supervivencia eran de otro tipo, menos académicas. Después de todo, para desviar el agua de una acequia un campesino no necesita el conocimiento explícito o *declarativo* del principio de Bernoulli, sino sólo el conocimiento implícito o *procedimental* de cómo se comporta el líquido y lo que hay que hacer para conducirlo. Esta asimetría entre las habilidades medidas por los

primeros tests y las habilidades que realmente tenía la población en aquel entonces explica que los resultados de CI fueran tan pobres. No es que nuestros antepasados fueran tontos, sino que eran inteligentes, pero de una manera tal que no necesitaban dominar las habilidades lógico-matemáticas y lingüísticas.

A medida que fueron pasando las décadas del siglo XX, la escolarización fue extendiéndose para formar obreros con destrezas intelectuales ajustadas a las nuevas necesidades. En un mundo cada vez menos rural y más industrializado, las competencias más apreciables pasaban a ser las de tipo lógico-matemático y las lingüísticas, justamente las examinadas por los tests de inteligencia. Por seguir con el ejemplo del agricultor, para cultivar los pequeños terrenos de los que se alimentaba la gente a principios de siglo bastaba con poseer un conocimiento procedimental del comportamiento del agua, mientras que las explotaciones de miles de hectáreas en las que se cultivan los alimentos en la actualidad requieren de sistemas de riego tan complejos que sólo pueden ser diseñados por profesionales que, como los ingenieros agrónomos y similares, conozcan el principio de Bernoulli. Y a su vez, para llegar a ser ingeniero agrónomo hay que haber superado sucesivas etapas de formación reglada cuyas pruebas de aptitud se centran principalmente en el examen de habilidades lógico-matemáticas y lingüísticas.

Por tanto, el efecto Flynn refleja el progresivo acercamiento de las habilidades reales de la población a las habilidades medidas por los tests. Si la psicometría hubiera sido fundada por agricultores y ganaderos, en vez de por hombres de ciencia, entonces los resultados de CI habrían ido bajando a medida que avanzaba el siglo XX. Los dominios y los ámbitos cambian de una cultura a otra, e incluso dentro de una misma cultura a través del tiempo, al igual que cambian los criterios de la belleza, y por tanto el otorgar un valor universal a las habilidades más apreciadas en la actualidad en Occidente es un error etnocéntrico. Los tests de inteligencia, dice Gardner, suelen cometer este error, dado que se centran en examinar las competencias lógico-matemáticas y lingüísticas como si estas dos por sí solas proporcionasen la medida universal de la inteligencia.

Distribuida

El tercer y último atributo característico de la inteligencia según Gardner es que es *distribuida*. Que la inteligencia es distribuida se refiere al hecho de que los seres humanos habitualmente trabajamos en grupo. El trabajo en grupo no sólo debe entenderse como la interacción directa con otros semejantes, sino también como la interacción indirecta que acontece cuando uno se sirve de los conocimientos y las tecnologías elaboradas por otros. Sobre este asunto, la *inversión del efecto Flynn* es un fenómeno esclarecedor. Tras la mencionada subida del CI durante todo el siglo XX, se observó que a partir de la década de 1980 las puntuaciones habían comenzado a disminuir. En el Reino Unido un estudio de 2009 reveló que entre 1980 y 2008 el CI de los estudiantes había caído aproximadamente dos puntos (Carr, 2010, p. 146). James Flynn explicó este fenómeno apelando al carácter distribuido de la inteligencia. Los tests de inteligencia, dice Flynn, siguen haciéndose mayormente en solitario y con lápiz y papel, exactamente igual que hace un siglo, mientras que los escolares en la actualidad resuelven sus problemas tanto académicos como cotidianos sirviéndose de tecnologías intelectuales como calculadoras y ordenadores. Cuando a un sujeto acostumbrado a utilizar estos dispositivos se le obliga a demostrar sus competencias en un test en el que no puede recurrir a ellos, es lógico que sus resultados sean peores que los un sujeto de la década de los 60, cuando las tecnologías intelectuales eran menos abundantes. Hacer hoy los tests como se hacían en el siglo pasado es como quitarle la pierna ortopédica a un cojo y pedirle que corra una maratón.

Así como nuestros antepasados no eran retrasados mentales, las nuevas generaciones no son menos inteligentes, sino que sus destrezas son distintas. La popularización de Internet, por ejemplo, ha dado lugar a un aumento de las competencias visuales y espaciales, al tiempo que ha disminuido el pensamiento profundo, entendiendo por tal aquel que comprende «la adquisición de conocimiento, el análisis inductivo, el pensamiento crítico, la imaginación y la reflexión» (Ibíd., p.

141). De la misma manera que Platón advertía en el *Fedro* de que el uso de la escritura afecta a la memoria, las nuevas tecnologías también están modelando la inteligencia. Que la inteligencia es distribuida significa que lo habitual es trabajar con los demás, ya sea de manera directa o bien indirecta a través de los conocimientos y las tecnologías que nos proporcionan.

5.3. El test de Turing

Dentro de un estudio sobre inteligencia artificial, en un capítulo dedicado a elucidar qué es la inteligencia no podía faltar una reflexión sobre el test de Turing. En 1954 Alan Turing se suicidó mordiendo una manzana empapada en cianuro tras haber sido declarado culpable de homosexualidad y condenado por un tribunal británico a elegir entre la cárcel o la castración química. Cuatro años antes de tan trágico final, el genio inglés dejó publicado un artículo que se convertiría en uno de los más comentados del siglo XX: *Computing machinery and intelligence*. Su propósito es averiguar si las máquinas inteligentes son técnicamente posibles.

En vez de abordar la cuestión de manera frontal definiendo lo que es una máquina y lo que es el pensamiento, que es justo lo que estamos haciendo nosotros, Turing opta por un planteamiento que él considera más productivo, y propone el *juego de imitación*, que es lo que en la actualidad conocemos como el *test de Turing*. El juego es simple: «Intervienen en él tres personas, un hombre (A), una mujer (B) y un interrogador (C). El interrogador permanece en una habitación, separado de los otros dos. El objeto del juego para el interrogador es determinar cuál de los otros dos es el hombre y cuál es la mujer. Los distingue mediante las letras X e Y, y al final del juego dice *"X es A e Y es B"* o *"X es B e Y es A"*. El interrogador puede formular preguntas de este tipo: *¿Podría decirme, X, la longitud de su pelo?* Supongamos que X es A, luego A ha de contestar. A trata de conseguir que X se equivoque al identificarla. [...] Para que el tono de la voz no ayude al interrogador, las respuestas deberían ser escritas, o mejor, escritas a máquina. La disposición ideal es un teletipo que comunique las dos

habitaciones. [...] El objeto del juego para el tercer jugador (B) es ayudar al interrogador. [...] Preguntamos ahora, *"¿Qué sucederá cuando una máquina se encargue del papel de A en este juego?"*» (Turing, 1950, p. 9). Según Turing, si la máquina es capaz de convencer al interrogador de que es un ser humano, entonces hay que concluir necesariamente que la máquina es inteligente. No se puede ir más allá y concluir que es un ser humano, dado que el interrogador no tiene acceso sensorial de ningún tipo a los sujetos A y B, pero sí se puede concluir que es inteligente, pues observa la conducta lingüística de ambos. Examinemos los compromisos epistemológicos que implica el otorgar tan decisiva importancia a la conducta y al lenguaje, comenzando por este último.

Turing no da detalles sobre cómo ha de ser la máquina del juego de imitación. No obstante, reconoce que la idea que tiene en mente es la de una computadora electrónica (Ibíd., p. 12). En tanto que computadora electrónica, se le suponen unas habilidades lógico-matemáticas extraordinarias. Por tanto, es razonable concluir que la máquina capaz de superar el juego de imitación posee al menos dos inteligencias: lógico-matemática y lingüística, justo las más valoradas por la tradición racionalista de Occidente. Dicha tradición tiene su origen en Platón y Sócrates, retoma su impulso en la Modernidad con Descartes y ya hemos visto que llega hasta nuestros días en forma de tests de inteligencia con un marcado sesgo etcnocéntrico según el cual las destrezas matemáticas y lingüísticas distinguen al sujeto inteligente. Turing hace suyos los dos criterios de inteligencia que vimos en Descartes al final del capítulo segundo. Éstos eran el lenguaje y la capacidad general de resolver problemas. Añadimos esta última capacidad, aunque Turing no la menciona explícitamente, porque el uso del lenguaje que ha de hacer la máquina para superar el juego de imitación no puede limitarse al cumplimiento de las reglas formales del idioma. Dado que el objetivo del interrogador es descubrir quién es la máquina, es de suponer que hará preguntas cuya respuesta correcta sólo esté al alcance de un ser humano normal, y los seres humanos normales tienen todos, como indica Descartes, una inteligencia general capaz de acometer problemas de todo tipo.

Ahora bien, hay una diferencia crucial entre Descartes y Turing que aleja a éste de la ortodoxia de la tradición racionalista. Y es que, mientras que para el filósofo francés la competencia lingüística es *condición necesaria* para la inteligencia, para Turing sólo es una *condición suficiente*. Según Descartes, hasta los mudos tienen su propio lenguaje de signos, y eso los distingue de los seres irracionales, no inteligentes (Descartes, *Discurso del método*, p. 94). En cambio, Turing no afirma que para ser inteligente sea necesario superar el juego de imitación y por tanto que sea necesario tener competencia lingüística, sino sólo que para superar el juego es necesario ser inteligente. Expresado formalmente esto se diría así: $\neg(I \rightarrow T) \wedge (T \rightarrow I)$, donde T es el enunciado "pasar el test" e I equivale a "ser inteligente". Si un sujeto pasa el juego de imitación, entonces se ha de concluir que es inteligente, pero si no lo pasa, no se concluye nada. Un chimpancé, por ejemplo, no superaría la prueba, dado que no tiene una competencia lingüística como la humana, pero no por ello debería concluirse que no es inteligente (Copeland, 1993, p. 79). Por tanto, Turing considera que la conducta lingüística indiscernible de la de un ser humano normal es por sí sola condición suficiente pero no necesaria de la inteligencia.

En cuanto a la conducta, Turing sostiene que es el único criterio que de hecho empleamos en realidad para decidir si alguien es inteligente. Si alguien o algo se comporta de manera inteligente, la convención social es concederle el atributo de la inteligencia (Turing, 1950, p. 29). En la novela de ciencia ficción *Solaris*, el escritor polaco Stanislav Lem narra las vivencias de un astronauta que viaja a un planeta cubierto totalmente por un enorme océano. El personaje al principio está convencido de que, al igual que los océanos de la Tierra, ése ha de ser también inerte, pero a medida que avanza la trama y observa la conducta de la gran masa líquida, se va viendo obligado a reconocer que se trata de un ser vivo inteligente. Hay un proverbio popular que dice algo así como: si parece vino, huele a vino y sabe a vino, entonces es vino. El mismo razonamiento se aplica a la inteligencia, según Turing. Si la conducta es la propia de un ser inteligente, entonces es inteligente. En un intento de explotar las deficiencias de este enfoque claramente conductista, Claude Shannon y John McCarthy

formularon en 1956, dos años después de la muerte de Turing, la siguiente objeción: «Una desventaja de la definición del pensamiento proporcionada por Turing es que es posible, en principio, diseñar una máquina con un conjunto completo de respuestas arbitrariamente elegidas para todas las posibles entradas sensoriales. Una máquina de ese tipo, ante cualquier entrada sensorial (incluyendo la historia pasada) solamente consultaría en su "diccionario" la respuesta correspondiente. Con el diccionario adecuado esa máquina pasaría el test de Turing, pero no refleja nuestro concepto intuitivo de lo que es la inteligencia» (Copeland, 2004, p. 437).

En opinión de Jack Copeland, el test de Turing se libraría de la objeción de Shannon y McCarthy si al *criterio de la conducta* en el que se basa se le añadiera el *criterio del diseño* (Copeland, 1993, p. 87). El criterio del diseño establecería que las respuestas deberían ser producidas por mecanismos similares a los que tienen lugar en el cerebro humano o bien por otros mecanismos distintos pero que, en cualquier caso, impliquen procesamiento de la información y no una simple búsqueda de correspondencias entre entradas y salidas en una base de datos. A nuestro juicio, sin embargo, el test de Turing no necesita ser modificado ni ampliado, en tanto que la objeción de Shannon y McCarthy incurre en la confusión de dos nociones de inteligencia. Por un lado está la inteligencia tal y como la entendemos de manera ordinaria, y por otro lado está la de los científicos. El criterio del diseño forma parte de esta última, pero no de la noción ordinaria. Si existiera una máquina como la descrita por Shannon y McCarthy, no sería calificada de inteligente por los científicos que hubieran examinado sus entrañas, pero sí por aquellas personas que interactuasen con ella de manera ordinaria, incluidos los propios científicos, sin abrirle la carcasa. SUPERPARRY, que es como Copeland bautiza a esa máquina en honor a un programa parecido a ELIZA, sería inteligente a ojos de aquellos que tratasen con él. Por otra parte cabe interpretar, dice Copeland, que Turing propone el juego de imitación como criterio válido no en cualquier mundo posible, sino en el mundo real, y en el mundo real no existen las máquinas como SUPERPARRY, por lo que la objeción de Shannon y McCarthy no sería aplicable (Copeland, 2004, p. 438).

El argumento de la sala china

Casi tan célebre como el propio test de Turing es la objeción contra él formulada por John Searle con su argumento de la sala china. Searle comienza por distinguir, como ya hicimos nosotros al comienzo del primer capítulo, entre IA fuerte e IA débil. Esta última, dice, es una realidad que resulta provechosa para la ciencia en tanto que su objetivo es construir modelos informáticos de ciertos procesos mentales que ayudan a mejorar los conocimientos de la psicología y la neurociencia. En cambio, rechaza la pretensión de la IA fuerte de construir inteligencias artificiales completas por considerar que en la actualidad, con la tecnología informática disponible, se trata de una empresa imposible. Para demostrar dicha imposibilidad, Searle revela las limitaciones formales de las computadoras electrónicas a través de su famosa metáfora de *la sala china*.

Supongamos, dice Searle, que un operario que sólo habla inglés se encuentra encerrado en una sala llena de papeles con símbolos escritos en chino. El operario dispone de un manual de reglas o instrucciones escrito en su idioma, el inglés, que le indica cómo debe relacionar unos símbolos chinos con otros, sin que sea necesario comprender lo que éstos significan. Por ejemplo, una regla podría ordenar: «Tome un signo garabato de la cesta número uno y colóquelo al lado de un signo garabís tomado de la cesta número dos» (Searle, 1990, p. 10). Fuera de la habitación hay una persona que sí habla chino y que le suministra al operario a través de una ranura pequeños grupos de símbolos chinos. El operario entonces manipula los símbolos entrantes siguiendo las reglas que figuran en el libro de instrucciones y devuelve al exterior otros conjuntos de símbolos. En esta analogía el libro de instrucciones hace las veces de programa informático, las personas que lo escribieron son los programadores y el operario es el procesador central de la computadora electrónica. Los cestos llenos de símbolos son la base de datos, los símbolos entrantes son las preguntas y los salientes son las respuestas. Supongamos ahora, continúa Searle, que el libro de instrucciones

estuviera escrito de tal manera que las respuestas producidas mediante el cumplimiento de sus indicaciones fueran indistinguibles de las que daría un nativo hablante del idioma chino. En tal caso, la sala china pasaría el test de Turing, y sin embargo el operario en su interior no habría entendido ni una sola palabra del intercambio de entradas y salidas de información lingüística. Por tanto, el test de Turing no es eficaz.

Para completar su argumento, Searle formula explícitamente los axiomas y las conclusiones contenidos en la metáfora de la sala china. El primer axioma es que los programas informáticos son formales, o lo que es lo mismo, sintácticos. Lo único que hacen las computadoras electrónicas, dice, es manipular símbolos siguiendo las reglas indicadas por un programa. Los símbolos por sí solos no tienen contenido semántico, en tanto que son manipulados sin referirse a ningún significado. Pueden tener significado para el ser humano que los observa, pero en cualquier caso la computadora los manipula como si no se refiriesen a nada, que es justo lo que hace el operario de la sala china. El segundo axioma es que la mente humana posee contenidos mentales, es decir, semánticos. De esta manera, cuando pensamos en un símbolo, como por ejemplo una palabra, nos viene a la mente el objeto al que se refiere. Y el tercer axioma afirma que la *sintaxis*, por sí misma, no es constitutiva ni suficiente para la *semántica*. La conclusión derivada de estos tres axiomas es, según Searle, que los programas informáticos ni son constitutivos de mentes, ni suficientes para ellas. La clave de todo el argumento «descansa en la distinción entre la manipulación formal de símbolos efectuada por el ordenador y los contenidos mentales que biológicamente existen en la mente, distinción que he abreviado –y espero no haber inducido a error a nadie– como distinción entre sintaxis y semántica» (Ibíd., p. 15).

La metáfora de la sala china, como se puede apreciar, desmitifica las propiedades de las computadoras electrónicas mediante el recordatorio histórico de que éstas, a pesar de lo sofisticado de la tecnología que las hace funcionar, no hacen nada distinto de lo que haría un computador humano. Los computadores humanos de las redes organizadas en el siglo XVIII por Gaspard De Prony y Nevil Maskelyne eran

capaces de producir, por ejemplo, informaciones astronómicas muy complejas sin saber absolutamente nada de astronomía. De igual modo, el operario de la sala china puede producir una conducta lingüística indistinguible de la de un nativo hablante de chino sin entender ni una sola palabra de ese idioma.

Objeciones al argumento de la sala china

Contra el argumento de la sala china se han propuesto numerosas objeciones que el propio Searle se ha encargado de recopilar y responder. De las que seis que vamos a examinar, la primera es la *objeción del sistema* (Searle, 1980, p. 288). Según ésta, es cierto que el operario no entiende el chino, pero el sistema entero, formado por la sala y todo lo que en ella hay, sí que lo entiende. La réplica de Searle consiste en suponer que el operario memorizase todos los documentos albergados en la sala, desde el libro de instrucciones hasta los cestos con símbolos. En tal caso, dice, es evidente que no habría nada en el sistema que no estuviera dentro del operario, y sin embargo, éste seguiría sin entender el idioma chino porque aún carecería de contenidos semánticos asociados a los símbolos. La segunda crítica es la *objeción del robot* (Ibíd., p. 293). Ésta propone imaginar un robot que, además de exhibir competencia lingüística, tuviera extremidades, articulaciones y aparatos sensoriales que le permitiesen percibir el mundo y moverse por él de manera semejante a como lo hace un ser humano. Habría que reconocer entonces, dicen los partidarios de la objeción, que un robot de este tipo tendría verdaderos estados mentales, con contenidos semánticos adquiridos gracias al contacto con la realidad. Searle responde pidiendo que imaginemos que, en el interior del cráneo metálico del robot, en vez de haber una computadora electrónica hubiese una sala china con un pequeño operario, un homúnculo. Éste recibiría símbolos chinos procedentes de los sistemas sensoriales y a su vez devolvería símbolos chinos hacia el aparato locomotor para producir el movimiento, pero seguiría sin entender nada de lo que sucede ahí afuera. De hecho, ni siquiera sabría lo que ocurre en el cuerpo que él gobierna.

El tercer argumento contra la sala china es el de la *objeción del simulador cerebral* (Ibíd., p. 294). Esta objeción cambia de programa de investigación de la IA, dado que se centra en la IA subsimbólica y deja la IA simbólica al margen. O dicho de otra manera, propone reproducir el cerebro, en lugar de la mente. Supongamos, dice esta tercera objeción, una computadora electrónica capaz de simular las secuencias de disparos neuronales que suceden en el cerebro de un nativo hablante de chino cuando éste entiende las historias en chino y responde a ellas. Podemos incluso suponer que la máquina realiza la simulación no de manera serial, como las máquinas von Neumann, sino en paralelo, que es como realmente el cerebro procesa la información. Deberíamos concluir entonces, dicen los defensores de la objeción, que una máquina como la descrita entendería el chino tan perfectamente como lo hace un ser humano hablante de ese idioma. Respecto a la *objeción de la combinación*, que es la cuarta, consiste, como su nombre sugiere, en una combinación de las anteriores, de manera que la máquina imaginaria sería un robot que dentro del cráneo, en vez de una IA simbólica, llevase una IA subsimbólica de simulación de redes neuronales como la recién descrita (Ibíd., p. 295). La réplica de Searle a ambas objeciones es que una red artificial de neuronas no produciría estados mentales porque, sencillamente, carecería de las propiedades causales del cerebro que producen dichos estados. Cuáles son esas propiedades discriminatorias es algo que Searle debería aclarar, pero no lo hace, razón por la cual esta réplica suya no resulta concluyente. En su defensa se limita a apelar a la analogía de que, así como la simulación de una tormenta no moja, la simulación de un cerebro no produce una mente. El agua tiene propiedades causales que mojan y el cerebro tiene propiedades causales biológicas que hacen emerger la mente.

Por último, las objeciones quinta y sexta son de carácter general, y por tanto aplicables a cualquier IA fuerte, ya sea simbólica, subsimbólica, híbrida o de cualquier otro tipo que se descubra en el futuro. En concreto, la quinta es la *objeción de la mente de los otros*, y consiste en defender la suficiencia del criterio de la conducta por sí solo para decidir si algo es inteligente, mientras que Searle estima necesario incluir el criterio del diseño (Ibíd., p. 297). Por su parte, la *objeción de las mansiones*

múltiples, que es la sexta, señala que los argumentos de Searle sólo son aplicables sobre las inteligencias artificiales tal y como las concebimos en la actualidad, pero no sobre las que pudieran ser inventadas en el futuro. A esto último el filósofo norteamericano responde que, en efecto, él está de acuerdo en que algún día quizás se diseñe una máquina que posea las propiedades causales del cerebro, pero dicha posibilidad no afecta a los argumentos que él sostiene contra la IA fuerte.

En lo que queda de esta sección vamos a rebatir dos de las réplicas de Searle. La primera y la segunda las abordaremos en capítulos posteriores cuando hablemos de la *IA abstracta* y la *IA situada* (Copeland, 2004, p. 439; Haugeland, 1996, p. 25), y la sexta no tiene discusión. Por tanto vamos a ocuparnos, en primer lugar, de rebatir la que apela a las misteriosas propiedades biológicas del cerebro que hacen emerger la mente, que es la clave de las objeciones tercera y cuarta, y a continuación, de la que considera necesario el criterio del diseño para distinguir a un sujeto inteligente. En cuanto a las propiedades biológicas del cerebro, Searle debería arriesgarse a concretar cuáles son, tal y como hace Penrose respecto de la conciencia. El matemático Roger Penrose, basándose en la mecánica cuántica, propone en su libro *Las sombras de la mente* que la conciencia se origina en unas estructuras citoesqueléticas de las neuronas denominadas *microtúbulos* (Searle, 1996, p. 665). Es una hipótesis muy falsable, y por tanto muy informativa, que es como deben ser las propuestas científicas. Por el contrario, decir que el cerebro posee propiedades biológicas desconocidas que son las que producen la mente es tanto como no decir nada, o peor todavía, es decir algo contrario a la evidencia empírica.

En el año 2003, el profesor Theodore Berger, de la Universidad del Sur de California, presentó un hipocampo artificial (Berger, 2005, p. 1). Desde entonces, el dispositivo ha sido probado con éxito en ratas, y por tanto no parece lejano el día en que un ser humano lo lleve alojado en el interior de su cráneo. La máquina de Berger parece hacer lo mismo que las neuronas de un hipocampo natural, que es una estructura muy importante para la inteligencia, tanto incluso como la corteza, pues hay quien la considera «la región suprema de la corteza cerebral y no una estructura

separada» (Hawkins & Blakeslee, 2004, p. 198). Partiendo de esta evidencia, cabe plantearle a Searle la siguiente cuestión de rampa escurridiza formulada originalmente por Zenon Pylyshyn (Crevier, 1993, p. 271): ¿Cuántas neuronas naturales deberían ser sustituidas por otras artificiales para que un sujeto dejase de tener mente? O quizás ni siquiera haya que especular tanto. Teniendo en cuenta que el hipocampo es la estructura principal implicada en la conversión de la memoria a corto plazo en memoria a largo plazo, según Searle los sujetos con la prótesis de Berger tendrían una mente a corto plazo pero no a largo, aunque se comportasen como si tuvieran una mente a largo plazo, es decir, basada en recuerdos. Todo un absurdo.

Contra el criterio del diseño

Respecto del criterio del diseño, opinamos al igual que Turing que desde el punto de vista de la experiencia ordinaria no es necesario. Las cosas son lo que son porque se comportan de determinada manera. Para averiguar, por ejemplo, si un trozo de metal es oro puro, de símbolo Au, los científicos comprueban características tales como que su densidad sea de 19.300 Kg/m^3, que su punto de fusión sea de 1.064 °C y que su conductividad eléctrica sea de $45{,}5x10^6$ S/m, entre otras muchas. Dado que a Searle le gustan los experimentos mentales, supongamos que se descubriese un nuevo elemento químico, y que en honor a él fuera bautizado como "johnsearlio", con símbolo químico "Js". Si el Js tuviese la misma apariencia que el Au y presentara en el laboratorio todas las propiedades que se utilizan para identificar el Au, entonces al johnsearlio se le llamaría "oro". En realidad no sería estrictamente oro, porque entre ambos elementos existirían diferencias a nivel subatómico. Ciertamente, esto es mucho suponer, pues las propiedades observables como las citadas surgen de las características subatómicas, pero en un experimento mental es aceptable concebir un elemento como el johnsearlio. Quizás en algunos experimentos científicos el Js no podría sustituir al Au debido a las sutiles diferencias subatómicas, pero en la vida ordinaria el johnsearlio y el oro serían la misma cosa.

Pues bien, lo mismo sucedería con una máquina que no tuviese representaciones mentales pero que se comportase de la misma manera que alguien que sí las tiene, o que en realidad fuera como SUPERPARRY y no procesara la información entrante, sino que la correlacionase con las respuestas adecuadas en una base de datos. Esa máquina, tal y como dice Turing y nosotros lo secundamos, sería calificada de inteligente. Es más, el propio Searle al tratar con ella por correo electrónico, o por cualquier otra vía que le impidiese la percepción sensorial, la consideraría inteligente. De hecho, si las inteligencias artificiales en sentido fuerte son posibles y en un futuro cercano se construye una, la primera tarea que debería serle ordenada por su creadores es cartearse con Searle y demás detractores del test de Turing. Para ser coherente con lo que escribió sobre la teoría extensional del significado de Hillary Putnam, Searle debería reconocer que el johnsearlio es oro y que SUPERPARRY es inteligente. Recordemos la polémica entre ambos filósofos.

Putnam comenzaba su artículo *The meaning of meaning* con un experimento mental (Putnam, 1973b, p. 700). Supongamos, dice, que en algún lugar del universo existe una Tierra Gemela en la cual todo es idéntico a la Tierra que conocemos. Todo lo que hay allí en cada instante es una copia exacta, partícula a partícula, de la Tierra, excepto por un detalle: que lo que allí llaman "agua" en realidad no está formado por H_2O, sino por otra fórmula química muy compleja que, para abreviar, puede ser denominada XYZ. Ambas sustancias se comportan de la misma manera, compartiendo el mismo punto de ebullición, grado de viscosidad y demás propiedades. Lo que Putnam sostiene es que cuando un sujeto de la Tierra y su homólogo de la Tierra Gemela piensan en el agua, en realidad no quieren decir (*mean*) lo mismo. En el momento en que sucede ese pensamiento sus cerebros son, como decimos, idénticos partícula a partícula, pero en opinión de Putnam no quieren decir lo mismo, porque uno se refiere a la sustancia H_2O mientras que el otro se refiere al XYZ.

Searle no está de acuerdo con Putnam, y en la sección *Are the meanings in head?* de su obra *Intentionality: An essay in the philosophy of mind* explica por qué. El precio a pagar por seguir las intuiciones de Putnam, dice, sería muy alto. Dado que son

muchas las cosas que tienen el agua como principal componente, si la sustancia de la Tierra Gemela no es agua, entonces el barro no es barro, la cerveza no es cerveza, la nieve no es nieve, los helados no son helados, la sopa no es sopa, y así sucesivamente. La manera más razonable de solucionar la disyuntiva, afirma, sería reconocer que hay dos tipos de agua: «Hasta el año 1750 "agua" significaba lo mismo y tenía la misma extensión en la Tierra y en la Tierra Gemela. Tras haber sido descubierto que hay dos composiciones químicas diferentes, una para la Tierra y otra para la Tierra Gemela, tendríamos que tomar una decisión. Podríamos *definir* el "agua" como H_2O, que es lo que hemos hecho; o bien podríamos simplemente decir que hay dos tipos de agua, y que el agua de la Tierra Gemela tiene una composición diferente a la del agua de la Tierra» (Searle, 1983, p. 203).

Para ser coherente, Searle debería asimismo conceder que SUPERPARRY es inteligente pero con otro tipo de inteligencia, en el mismo sentido en que XYZ es otro tipo de agua. Por el contrario, si no se concediera que SUPERPARRY es inteligente, entonces tendríamos que inventar palabras nuevas para referirnos a sus conductas. Así, suponiendo que SUPERPARRY exhibiera una conducta indistinguible de la de un ser humano con un grado normal de inteligencia, su amabilidad no sería amabilidad, su locuacidad no sería locuacidad y sus reflexiones no serían reflexiones. SUPERPARRY sería como un judío en la Alemania nazi. Sin necesidad de suponer la existencia de la Tierra Gemela, si en el propia Tierra se descubriera que el agua salada es en realidad XYZ a causa de un fenómeno hasta ese momento desconocido, entonces el agua salada seguiría siendo agua, diría Searle. De la misma manera, Searle debería conceder el atributo de la inteligencia a un sujeto que se comporta de manera tan inteligente como un ser humano. De hecho, no nos cabe la menor duda de que lo haría, tal y como le sucedió a la secretaria de Weizenbaum.

Como vimos en el primer capítulo, en 1966 Joseph Weizenbaum publicó ELIZA, un programa informático de conversación basado en scripts, como el SAM de Roger Schank, que enseguida se hizo muy popular en los círculos universitarios. En *Computer power and human reason* Weizenbaum cuenta que su secretaria pasaba largos ratos

charlando con ELIZA mediante un terminal electrónico (Weizenbaum, 1976, p. 17). Como la situación le parecía aberrante, le explicó a la mujer que en realidad al otro lado del terminal no había nadie, sino que ELIZA era un programa de ordenador incapaz de comprender lo que se le decía, o lo que es lo mismo, no tenía representaciones mentales. A pesar de que se trataba de una persona con un nivel cultural medio, la mujer siguió dedicando tiempo a hablar con ELIZA a escondidas. La máquina le proporcionaba la reconfortante sensación de que había alguien escuchando sus problemas. Su conducta estaba justificada porque las cosas son lo que son en función de cómo se comportan. Por qué se comportan de una forma y no de otra es cuestión para los científicos, así como es cuestión para los científicos el diferenciar entre el oro y el johnsearlio. Si las conductas cotidianas de tratar al johnsearlio como si fuera oro y de tratar a una computadora como si fuera inteligente resultan exitosas, entonces son válidas, porque el éxito es el criterio último de decisión de todas las teorías, ya sean científicas o precientíficas.

La importancia de la inteligencia lingüística

Como demuestra la mencionada objeción del chimpancé (Copeland, 1993, p. 79), el test de Turing es un criterio suficiente aunque no necesario para decidir si un sujeto es inteligente, pues un mono, o cualquier otro animal, no pasaría el test, y sin embargo tiene un cierto grado de inteligencia. Ahora bien, dado que nuestro interés se centra en la reproducción artificial de la inteligencia humana, y no de la inteligencia en general entendida como un atributo presente en otras especies, consideramos que la conducta lingüística evaluada por el test de Turing es un criterio necesario y suficiente para la inteligencia humana más distintiva, la lingüística, en tanto que todo ser humano es capaz de comunicarse mediante algún tipo de lenguaje.

Ser humano implica la capacidad de comunicación lingüística con otros seres humanos, sea de forma oral o gestual. Gardner observa que: «El don del lenguaje es universal, y su desarrollo en los niños es sorprendentemente similar en todas las

culturas. Incluso en el caso de personas sordas a las que no se ha enseñado explícitamente un lenguaje por signos, a menudo de niños "inventan" su propio lenguaje manual y lo usan subrepticiamente. Vemos así que una inteligencia puede operar independientemente de una cierta modalidad de salida o de un determinado canal de salida» (Gardner, 1993, p. 44). Lo mismo dice Descartes: «Hasta los hombres que han nacido sordos y mudos se hacen entender por los que viven con ellos y tienen tiempo de aprender su lenguaje» (Descartes, *Discurso del método*, p. 94). Un ciego podría pasar el test de Turing utilizando un terminal de braille, al igual que un analfabeto sordo y mudo podría hacerlo con la ayuda de un intérprete que leyese las preguntas del interrogador y transcribiera sus respuestas. Si una máquina superase el test de Turing, entonces demostraría ser poseedora de una de las dos inteligencias que Descartes señala como exclusivas del ser humano: la lingüística. La otra, que es la inteligencia general para resolver todo tipo de problemas (Ibíd., p. 93), también puede ser, a nuestro juicio, evaluada en cierta medida mediante el test de Turing.

Según la teoría IM, que nosotros suscribimos, el espectro de todas las tareas realizables por la mente se agota en siete inteligencias modulares. Dado que esas siete inteligencias, como apunta Gardner, son susceptibles de ser codificadas en sistemas de símbolos, y dado que cualquier sistema de símbolos es descriptible mediante el lenguaje ordinario gracias a la flexibilidad de éste, el test de Turing también es un criterio suficiente para decidir que un sujeto posee, al menos parcialmente, aquellas otras inteligencias cuyos sistemas simbólicos domine en la conversación. John Haugeland respalda esta observación: «Hablar no es una mera habilidad entre otras, sino además, y esencialmente, la habilidad de *expresar* inteligentemente muchas (quizás todas) otras habilidades de la inteligencia. Y sin *poseer* esas habilidades de hecho, o al menos en cierto grado, uno no puede hablar de forma inteligente sobre ellas. Por eso el test de Turing es tan irresistible y poderoso» (Haugeland, 1996, p. 4). Por ejemplo, si un sujeto sometido al test de Turing mostrase un gran conocimiento sobre danza, entonces debería reconocérsele una cierta inteligencia espacial y cinético-corporal, al menos a nivel declarativo, aun a pesar de no poder demostrarlo a

nivel procedimental por no tener piernas, pues también hay bailarines que, por achaques de la edad o por accidente, no pueden bailar y sin embargo se los escucha atentamente cuando hablan sobre el tema. Ahí están los críticos de cine que no han filmado ni un bautizo, y sin embargo se les concede que saben de cine porque en sus crónicas manejan con soltura términos como *travelling*, *flashback* y *storyboard*, y sin hablar inglés muchos de ellos.

A propósito del idioma inglés: la palabra *dumb* significa "tonto" y "mudo". La razón de esta polisemia, hoy en día socialmente censurada, es que antiguamente en Inglaterra a los mudos se les llamaba tontos por no poder hablar. Recordemos que "golem" en hebreo significa "tonto". El golem del rabino Low era tonto: no tenía inteligencia, no sabía hablar. Lo mismo pensaríamos de una IA que no tuviera inteligencia lingüística, ni para entender ni para hacerse entender: que es tonta. Podría tener todas las demás inteligencias: musical, cinético-corporal, lógico-matemática, espacial, interpersonal e intrapersonal, pero sin la lingüística no diríamos que su inteligencia está a la altura de la de un ser humano. La consideraríamos más cerca del resto de los animales que de nosotros. Como dice Wittgenstein: «El hombre posee la capacidad de construir lenguajes en los que cualquier sentido resulte expresable, sin tener la menor idea de cómo y qué significa cada palabra. Al igual que se habla sin saber cómo se producen los diferentes sonidos. El lenguaje ordinario es una parte del organismo humano y no menos complicado que éste» (Wittgenstein, 1921, §4.002). Todos los animales compartimos el mismo mundo físico, pero al social de los seres humanos sólo se accede a través del lenguaje. Las competencias cinético-corporales, espaciales, interpersonales e intrapersonales se encuentran repartidas por el reino animal. En cuanto a las musicales y lógico-matemáticas, hace tiempo que se inventaron las máquinas con habilidades creativas, como componer música (Kurzweil, 1999, p. 160) y descubrir teoremas matemáticos como hizo el Logic Theorist de Newell y Simon del que hablaremos en el siguiente capítulo (Crevier, 1993, p. 46), pero socialmente no son consideradas auténticas inteligencias artificiales en sentido fuerte, porque es la inteligencia lingüística la que produce el mundo social y da acceso a él.

En consecuencia, una máquina poseedora de todas las inteligencias menos la lingüística estaría excluida de nuestro mundo más propio. Sería inteligente, sin duda, pero estaría demasiado lejos de ser *casi* tan inteligente como un ser humano, y por tanto no sería una auténtica IA fuerte. En un contexto cultural logocéntrico como el nuestro la inteligencia lingüística, para bien o para mal, está sobrevalorada, y es justamente ese valor excesivo lo que le confiere su carácter decisivo en el ámbito de la IA. Estamos de acuerdo con Gardner y Schank en que no se debería juzgar la inteligencia en general de un ser humano por su destreza lingüística, pero de hecho es así como se juzga. Quizás tampoco se debiera juzgar del mismo modo la inteligencia en general de las máquinas, pero también es así como se juzga. Prueba de ello es que en la Historia de la IA, como veremos en el capítulo siguiente, los mayores esfuerzos se han dedicado a la creación de máquinas capaces de entender y hablar el lenguaje natural (Dreyfus, 1992, p. 91). Por tanto, en lo que respecta a la IA, el test de Turing, por la importancia que concede a la conducta lingüística y por la importancia de ésta en nuestra cultura, es la piedra de toque para una verdadera IA fuerte.

Dentro de nuestra aceptación del test de Turing, estimamos, no obstante, que puede ser mejorado añadiéndole una escala del tipo de la antes mencionada de los tres niveles de comprensión distinguidos por Schank: tener sentido (*making sense*), comprensión cognitiva (*cognitive understanding*) y empatía completa (*complete empathy*) (Schank, 1986, p. 124). El sujeto de la otra habitación, ya sea un hombre o una computadora electrónica, mostrará que las entradas y salidas de información tienen sentido para él si es capaz de «unir eventos rellenando los huecos para asegurarse de que existe una cadena de causalidad sólida» (Ibíd., p. 128). Estará en el escalón superior de la comprensión cognitiva si es capaz de «recuperarse de expectativas incumplidas mediante la rememoración de experiencias anteriores similares, y es capaz de aprender a partir de la comparación del fallo actual con dichas experiencias previas» (Ibíd., p. 128). Y tendrá el grado más alto de inteligencia si es capaz de formar una representación acertada de la mente del interrogador que le permita ponerse en su lugar y, eventualmente, incluso adelantarse a sus palabras

(Ibíd., p. 129). A éste último nivel, caracterizado por la capacidad de anticipación, es al que se refiere Hawkins cuando dice que: «Si la habitación china de Searle contuviera un sistema de memoria similar (al del cerebro humano) que pudiera realizar predicciones sobre qué caracteres chinos aparecerían a continuación y qué pasaría después en el relato, estaríamos en situación de garantizar que la habitación entendía chino y comprendía el relato» (Hawkins & Blakeslee, 2004, p. 126). Ahora bien, si es posible lograr esto con un dispositivo que sólo es capaz de efectuar algoritmos, valga decir como una sala china o como una computadora electrónica, es la próxima cuestión de la que nos vamos a ocupar.

Resumen

Mente y cerebro operan de una manera circular, de lo particular a lo universal y viceversa, que no se ajusta al funcionamiento de las computadoras electrónicas. Ciertamente, las computadoras funcionan en bucles, pero esos bucles no son como los del círculo hermenéutico, porque en cada ciclo exigen la precisión absoluta propia de los lenguajes formales, mientras que en el círculo hermenéutico, tanto a nivel cerebral como mental, los ciclos van afinando las representaciones de manera difusa, imprecisa y aproximativa. Es paradójico que desde los tiempos de Pascal los ingenieros más brillantes pasasen siglos estrujándose los sesos para construir mecanismos de la máxima exactitud, hasta que finalmente lo lograron con las computadoras electrónicas, y desde entonces llevan décadas intentando lo contrario: utilizar esas máquinas, que de tan exactas son casi digitales, para emular el funcionamiento de un órgano, el cerebro, y su epifenómeno, la mente, cuya característica más destacada es justamente la imprecisión, la vaguedad, la tolerancia al fallo.

En el capítulo séptimo examinaremos si es técnicamente posible programar computadoras, tal y como las describimos en el capítulo tercero, para producir inteligencia, tal y como hemos caracterizado este concepto a través del examen de las teorías de Jeff Hawkins y Roger Schank. Veremos si las tesis de la neurociencia, como

las de Hawkins, son realizables en forma de IA subsimbólica, y si las tesis de la psicología cognitiva, como las de Schank, son realizables en forma de IA simbólica. En cuanto a la teoría IM de Gardner, nos ha servido para entender la inteligencia desde un punto de vista integral que contrasta con los planteamientos eliminativistas de Hawkins, que es fisicalista, y de Schank, que es mentalista.

6. Historia de la IA

«Me llamo Robinette Broadhead, pese a lo cual soy varón. A mi analista (a quien doy el nombre de Sigfrid von Schrink, aunque no se llama así; carece de nombre por ser una máquina) le hace mucha gracia este hecho. […] –Rob, hoy no estás cooperando mucho –dice Sigfrid a través del pequeño altavoz que hay en el extremo superior de la alfombra. A veces utiliza un muñeco de aspecto muy real, que está sentado en un sillón, da golpecitos con un lápiz y me dedica una rápida sonrisa de vez en cuando. […] A veces intento esto con él, diciendo alguna verdad dolorosa con el tono de quien pide otro ponche de ron al camarero de una fiesta. Lo hago cuando quiero esquivar su ataque. No creo que surja efecto. Sigfrid tiene muchos circuitos Heechee en su interior. Es mucho mejor que las máquinas del instituto al que me enviaron durante mi episodio. Observa continuamente todos mis parámetros físicos: conductividad cutánea, pulso, actitud de ondas beta, en fin, de todo. Obtiene indicaciones de las correas que me sujetan sobre la alfombra, acerca de la violencia con que me retuerzo. Mide el volumen de mi voz y lee sus matices en el espectro. Y también conoce el significado de las palabras. [...]

»Si Sigfrid fuese una persona de carne y hueso, no podría resistir todos los problemas que descargan en él. Para empezar, él ya tendría sus propios problemas. Tendría los míos, porque así es como yo me libraría de ellos, descargándolos en él. También tendría los de aquellos que, como yo, ocupan este diván; y él descargaría todo esto, porque tendría que hacerlo, en el hombre que estuviera por encima de él, en el que le psicoanalizara a *él*, y así sucesivamente hasta llegar a... ¿qué? ¿El fantasma de Sigmund Freud? Pero Sigfrid no es real. Es una máquina. No puede sentir el dolor. Así pues, ¿adónde *van* a todo ese dolor y ese cieno?» (Pohl, 1977).

En estos fragmentos de su novela de ciencia ficción *Pórtico*, el escritor Frederik Pohl fantasea con la posibilidad de que en el futuro existan inteligencias artificiales en sentido fuerte que, como Sigfrid, sean capaces de reemplazar a los seres humanos en las tareas que nos son más propias y que requieren de un alto nivel de inteligencia, como la psicoterapia. Sigfrid sería una versión sofisticada de DOCTOR, el programa desarrollado por Kenneth Colby en base al ELIZA de Joseph Weizenbaum, y que éste repudió por considerarlo inmoral. Sobre la moralidad de la IA hablaremos en el capítulo octavo, dedicado a las condiciones de posibilidad sociales de esta tecnología, y sobre las condiciones de posibilidad técnicas, en el séptimo. El presente vamos a dedicarlo a examinar la Historia de la IA desde su fundación en los años 50 hasta la actualidad. Lo haremos siguiendo el hilo conductor del extraordinario libro de Daniel Crevier *The tumultuous history of the research for artificial intelligence*, y aprovecharemos el recorrido para exponer los conceptos técnicos más importantes de esta disciplina, de manera que también descubriremos cómo se las han ingeniado los informáticos para intentar crear máquinas pensantes.

6.1. Historia de la IA

En el verano de 1956, un mes antes de que en Massachusetts tuviera lugar el simposio considerado como el momento fundacional del paradigma cognitivista, en el estado vecino de New Hampshire se celebró la *conferencia de Dartmouth*, el momento fundacional de la IA. Dos jóvenes brillantes, el matemático de Princeton John McCarthy y el ingeniero del MIT Marvin Minsky, habían convencido a la Fundación Rockefeller para que cubriese los gastos de un taller de verano de dos meses de duración en el campus del Dartmouth College, donde se trataría el asunto de las máquinas pensantes. El visto bueno para obtener la financiación lo consiguieron gracias al aval de sus colegas Nathaniel Rochester, diseñador de la primera computadora electrónica de IBM producida en serie, la 701, y Claude Shannon, el creador de la teoría de la información, de la que ya hablamos en el capítulo cuarto, y

que conocía a McCarthy y Minsky porque ambos habían trabajado a sus órdenes durante el verano de 1953 en Bell Labs, los laboratorios de investigación de la compañía telefónica AT&T. Junto con estos cuatro, los otros seis asistentes a la conferencia de Dartmouth fueron Ray Solomonoff, Oliver Selfridge, Trenchard More y Arthur Samuel. Y nombramos aparte a los dos últimos asistentes, Herbert Simon y Allen Newell, porque fueron los únicos que se presentaron allí con algo más que ideas. Llegaron nada menos que con una IA bajo el brazo, quizás la primera de la Historia: el *Logic Theorist* (Lógico Teórico).

Con la ayuda del programador de sistemas John Clifford Shaw, Newell y Simon habían escrito los algoritmos del Logic Theorist un año antes de la conferencia. Como en aquella época las computadoras electrónicas eran bienes escasos y de difícil acceso, para comprobar el funcionamiento del programa emplearon la estrategia de la sala china de Searle y de los computadores humanos de De Prony y Maskeline, y lo dieron a ejecutar a un conjunto de personas que ignoraban cuál sería la consecuencia global de todos aquellos cálculos que unos algoritmos escritos en papel les ordenaban hacer. Gracias a que Shaw trabajaba en la RAND Corporation, una organización de investigaciones científicas con fines militares, al año siguiente, en 1956, pudieron ejecutarlo en una computadora electrónica, donde demostró funcionar a la perfección. La habilidad del Logic Theorist era demostrar teoremas de lógica extraídos de la obra *Principia mathematica* de Bertrand Russell y Alfred North Whitehead. De los primeros 52 teoremas del capítulo segundo de dicha obra, el Logic Theorist fue capaz de probar 38, con el mérito adicional de que su demostración del teorema número 2.85 era más elegante que la ofrecida por Russell y Whitehead (Crevier, 1993, p. 46).

El diseño del Logic Theorist estaba fuertemente influido por las ideas del emergente paradigma cognitivista. Sus creadores intentaron que fuese un programa de IA simbólica que realizase las demostraciones lógicas operando con símbolos de la misma manera que lo haría la mente de un ser humano (Ibíd., p. 44). Este enfoque *realista* se denomina *IA humana* (*human AI*) o *IA de modo teórico*, también llamada *simulación cognitiva*, frente al planteamiento *instrumental* de la *IA ajena* (*alien AI*) o *IA*

de modo operativo, que pretende crear inteligencias artificiales realizando cualesquiera operaciones, aunque no se parezcan a las que realizaría un ser humano, con tal de que el resultado sea una conducta inteligente (Copeland, 1993, p. 54). Alan Turing caracteriza ambas posturas con una metáfora que deja clara su opinión: si se dedicaran a construir automóviles, dice, los investigadores de la IA humana intentarían diseñar artefactos que se desplazasen utilizando piernas mecánicas, mientras que los de la IA ajena podrían permitirse explorar otras opciones, como las ruedas (Turing, 1948, p. 420). Partidarios de la IA humana, los creadores del Logic Theorist se esforzaron por replicar el carácter autoasociativo de la memoria humana y las operaciones de construcción y destrucción de estructuras simbólicas que, de acuerdo al cognitivismo, se supone que acontecen en la mente (Crevier, 1993, p. 47).

Por ser los únicos que acudieron a la cita con una IA terminada y operativa, Newell y Simon se mostraron altivos hacia el resto de participantes de la conferencia de Dartmouth. Pero a pesar de esa distancia hubo consenso en definir lo que luego daría a conocerse como la *hipótesis del sistema de símbolos físico* (*physical symbol system hypothesis*), o simplemente *hipótesis del sistema de símbolos* (HSS). En palabras de Newell y Simon, la HSS es a la IA simbólica lo que la doctrina celular a la biología o la teoría de las placas tectónicas a la geología (Newell & Simon, 1975, p. 38), es decir, el supuesto nuclear que debe ser defendido a toda costa mediante la heurística negativa. Una de sus posibles formulaciones es la que establece que: «Todos los aspectos del aprendizaje o cualquiera otra característica de la inteligencia puede, en principio, ser descrita con la precisión necesaria para que pueda construirse una máquina que la simule» (Crevier, 1993, p. 48). Otra formulación es la que asevera que: «Un sistema de símbolos físico tiene las capacidades necesarias y suficientes para la acción inteligente general» (Newell & Simon, 1975, p. 41). Un *sistema de símbolos* es lo que en el capítulo tercero definimos como un sistema formal, es decir, «una serie de proposiciones dispuestas en tal forma, que de algunas de estas proposiciones, llamadas *axiomas*, se derivan otras proposiciones con ayuda de ciertas reglas de inferencia» (Ferrater, 1965a, p. 689).

El adjetivo "físico" denota que se trata de la realización material de un sistema formal, y no un mero sistema formal que sólo existe a modo de idea. Decir que es físico es afirmar que dicho sistema obedece a las leyes de la física, en tanto que es realizable por medio de la ingeniería (Newell & Simon, 1975, p. 40). Así pues, un sistema de símbolos físico es un sistema formal ejecutado por una máquina real, típicamente una computadora electrónica. No obstante, no se adquiere el compromiso de que dicha máquina deba ser una computadora electrónica, pues por ejemplo el cerebro sería otro tipo de máquina que, supuestamente, también produce la inteligencia ejecutando un sistema de símbolos. «La inteligencia es nada más que la capacidad de procesar símbolos. Reside en un reino diferente al del hardware que la soporta, lo trasciende, y puede adoptar diferentes formas físicas» (Crevier, 1993, p. 48). Se observa claramente que con la HSS la IA suscribió en sus orígenes uno de los dos supuestos nucleares del paradigma cognitivista, a saber: la *tesis internalista*, según la cual existen representaciones mentales y es posible estudiarlas al margen de las estructuras materiales de las que surgen, ya sean éstas el cerebro en el caso de los seres humanos o el hardware en el de las computadoras electrónicas.

Existe una variante de la hipótesis del sistema de símbolos conocida como la *hipótesis fuerte del sistema de símbolos* (*strong physical symbol system hypothesis*), o HFSS, según la cual no es que un sistema de símbolos físico tenga las capacidades necesarias y suficientes para la acción inteligente general, sino que dichas capacidades son *exclusivas* de los sistemas de símbolos físicos. Desde este punto de vista radical de la IA simbólica, la posibilidad técnica de la IA fuerte estaría garantizada por principio, dado que el cerebro humano mismo sería un tipo de computadora, y la mente, el software ejecutado (Copeland, 1993, p. 273). Esto es justo lo que dice el otro supuesto nuclear del cognitivismo: la *metáfora computacional*, que es la analogía que define la mente como un procesador de información semejante a una computadora electrónica. Hoy en día son pocos los científicos que todavía defienden la HFSS, pues gracias al progresivo descubrimiento de la paradoja computacional (Gardner, 1985, p. 414) en la actualidad sabemos con certeza que la mente no opera como una computadora

electrónica. Sin embargo, es importante recordar la HFSS, ya que en los inicios del cognitivismo y de la IA tuvo partidarios tan destacados como los propios Newell y Simon (Franklin, 1995, p. 102).

Durante las dos décadas posteriores a la conferencia de Dartmouth, todos los avances significativos en la recién fundada disciplina de la IA fueron realizados por los diez asistentes o por sus estudiantes (Crevier, 1993, p. 49). Los dos grandes temas en los que centraron sus esfuerzos, y que todavía hoy siguen siendo problemáticos, fueron la heurística y el aprendizaje. El asunto del aprendizaje de las máquinas ya lo hemos tratado en el capítulo anterior, así que aquí vamos a referirnos sólo a la *heurística*, término acuñado en la primera mitad del siglo XX por el matemático George Polya para designar las *reglas generales* (*rules of thumb*) que permiten resolver problemas (Ibíd., p. 43). Para comprender en qué consiste la heurística, veamos un ejemplo de IA que no requiere heurística y otro que sí la necesita. El primer caso es el de una IA jugadora de tres en raya, y el segundo, el de una IA jugadora de ajedrez.

Heurística

Supongamos que queremos programar una computadora electrónica para jugar al tres en raya, también conocido como *tic tac toe*. El primer jugador que coloca una ficha, sea el humano o sea la máquina, ha de elegir entre 9 movimientos legales, tantos como casillas hay libres, pues en este juego pocas reglas hay aparte de la de que las fichas han de ponerse en casillas desocupadas. A continuación, el segundo jugador tiene que elegir entre 8 movimientos legales. Cuando vuelve a tocarle el turno al primer jugador, dispone de 7 movimientos posibles, y así sucesivamente hasta que todas las casillas están ocupadas o uno de los dos jugadores ha conseguido colocar tres de sus fichas en línea y por tanto el juego ha terminado. Si hiciéramos un esquema de todos los movimientos legales en una partida de tres en raya, tendría la forma de un árbol, llamado *árbol de decisión* (*decision tree*), en el que cada movimiento posible sería un nudo o *nodo* (*node*), del que crecen *ramas* (*branches*) que conducen a otros

nodos de nivel inferior. De esta manera, en el primer turno habría 9 nodos de cada uno de los cuales salen 8 ramas, en el segundo turno habría 9x8 nodos con 7 ramas cada uno, en el tercero, 9x8x7 nodos con 6 ramas cada uno, y así sucesivamente. Cuantas más ramas salen de un nodo, mayor es su *anchura* (*width*), y cuantos más niveles de nodos tiene por debajo, mayor es su *profundidad* (*depth*).

Utilizando el algoritmo *Minimax*, inventado por John von Neumann en un teorema homónimo de 1926, una IA para jugar al tres en raya consistiría en dos grandes módulos o funciones: una *función de movimiento* y una *función de evaluación* (Winston, 1981, p. 117). La función de movimiento proporciona a la función de evaluación una representación de los movimientos posibles, por ejemplo en forma de árbol mediante un proceso como el que acabamos de describir. Por su parte, la función de evaluación es la encargada de asignar un valor numérico a cada nodo. Asigna un número positivo a las situaciones del tablero que son favorables para el denominado *sujeto maximiza*nte, que en este caso sería la computadora porque es el jugador que está ejecutando el algoritmo Minimax, mientras que asigna un número negativo a las que son desfavorables a dicho sujeto, o lo que es lo mismo, son favorables para el oponente. Cuanto más positivo sea el número asignado a un nodo, mejor es la situación representada por él, y cuanto más negativo, peor.

La asignación de valores se realiza de abajo a arriba: comenzando por los nodos del nivel más bajo, que son los que están más alejados en el futuro, y trasladando sus valores hacia el presente. Por ejemplo, el valor de un nodo X situado en el turno 5 vendrá determinado por los valores de los nodos del turno 6 que surgen de él. Supongamos que la máquina mueve primero, y por tanto le tocará mover siempre en los turnos impares. Dado que en el turno 6 quien mueve es el humano, es de suponer que, de todos los nodos del turno 6 disponibles desde X, elegirá el de valor más bajo, o *mínimo*, pues es el que más le conviene. Si los nodos del turno 6 que salen de X tienen valores respectivos de 2, 8, -1, 7 y -5, el valor de X será, por tanto, -5. A su vez, X es uno de los nodos del turno 5 en los que se ramifica el nodo Y del turno 4. El valor de Y estará determinado por dichos nodos de nivel 5. Como el turno 5 es impar y en los

turnos impares mueve la máquina, ésta elegirá el nodo de valor más alto, o *máximo*. A su vez, Y es uno de los nodos del turno 4 en los que se ramifica el nodo Z del turno 3. El valor de Z estará determinado por dichos nodos de nivel 4. Como el turno 4 es par y en los turnos pares mueve el humano, elegirá el nodo de valor más bajo, o *mínimo*. Y así sucesivamente hasta determinar el valor del nodo N del turno 1. De la alternancia en la asignación de los valores máximos y mínimos procede el nombre del algoritmo: Minimax. Una vez determinados todos los valores de todos los nodos gracias a la función de evaluación, la máquina sabe perfectamente qué nodo o movimiento ha de elegir cuando es su turno: siempre el de valor más alto.

Supongamos ahora que quisiéramos utilizar el algoritmo Minimax tal y como lo hemos descrito para programar una IA jugadora de ajedrez. Tendría que calcular, por tanto, los valores de todos los nodos posibles, que es lo que hemos hecho en el tres en raya. En el ajedrez hay 20 movimientos posibles para las blancas durante, más o menos, sus 5 primeros turnos, y otros 20 para las negras. A partir del sexto turno, la cifra es de aproximadamente 35 movimientos posibles. Dado que una partida media de ajedrez consta de unas 40 jugadas por jugador, o lo que es lo mismo, 80 turnos o niveles de nodos, el número total de nodos o situaciones de tablero que se pueden dar es de $(20x20)^5x(30x30)^{35}$. El resultado de esta operación es $2x10^{116}$ nodos, una cifra que excede abismalmente las capacidades de cualquier computadora electrónica. Y si contemplamos las partidas que se extienden más allá del turno 80, el número asciende a $10^{18.900}$. Para mesurar estas magnitudes, piénsese que el número de átomos que hay en el universo se estima en torno a 10^{80}. Por tanto, no es posible programar una computadora para que juegue al ajedrez empleando la estrategia de calcular por *fuerza bruta* (*brute force*) todas las posibilidades como en el tres en raya, un juego en el que el número de nodos es de sólo 255.168.

Las estrategias empleadas para crear inteligencias artificiales que tienen que habérselas con un número insondable de posibilidades se denominan *estrategias heurísticas*, las cuales se definen como «procedimientos que, aunque no garantizan dar la respuesta correcta en todas las ocasiones, son fiables la mayor parte del

tiempo» (Haugeland, 1996, p. 13). En el caso del ajedrez, una estrategia heurística sería la *poda alfa-beta* (*alpha-beta prune*). Ésta consiste básicamente en restringir la anchura y profundidad de la parcela del árbol que debe ser examinada mediante la aplicación de criterios falibles. Por ejemplo, dada una situación en la que es posible liquidar a la reina rival con un simple movimiento de peón o bien exponer al rey propio en medio del tablero, los nodos que se siguen de esta última posibilidad no serán evaluados por considerarse que comparativamente con la primera no conducirán a un desenlace tan positivo. Así se consigue reducir el número de nodos a examen. Es una estrategia falible, pues puede que el movimiento del rey, aunque parece disparatado, conduzca a resultados probabilísticamente más favorables, pero dada la finitud de la capacidad de computación, hay que descartar opciones, y es razonable introducir en el código de la IA la regla general (*rule of thumb*) de que exponer al rey en el centro del tablero es una mala decisión que sólo debe ser contemplada cuando las alternativas sean igual de malas o peores.

Por supuesto, las estrategias heurísticas consisten en algoritmos, pues todo cuanto pueden hacer las computadoras electrónicas es ejecutar algoritmos. Pero, desde un punto de vista más amplio, se considera que no son procedimientos algorítmicos, sino heurísticos, en tanto que, a diferencia de los algoritmos, no siempre conducen al fin deseado. La regla general de no exponer al rey es formalizable algorítmicamente, pero se trata de una estrategia heurística porque es falible. Es una cuestión de perspectiva (Ibíd., p. 14).

Primeros programas

Tras el éxito del Logic Theorist, Newell y Simon siguieron explorando en la línea de la IA humana. Su siguiente creación fue el *General Problem Solver* (resolutor general de problemas), o GPS, un programa ejecutado por vez primera en 1957 y que pretendía simular los procesos cognitivos humanos que sirven para solucionar cualquier tipo de problema. La heurística empleada por el GPS, y que supuestamente

empleamos también los seres humanos, era el *análisis de medios-fines* (*means-ends analysis*), el cual consiste en detectar las diferencias entre el estado de cosas actual y la meta deseada, y reducirlas progresivamente (Crevier, 1993, p. 53). Aplicando el análisis de medios-fines, si un mono quisiera alcanzar un plátano situado en un lugar muy alto, procedería buscando medios para acortar la distancia, tales como colocar una silla debajo, intentar alcanzar la fruta con un palo, o combinar ambas acciones.

En realidad, como señala Roger Schank, ésta no es la manera en que resolvemos la mayoría de los problemas (Schank, 1999, p. 217). En vez de razonar partiendo de cero, lo que hacemos tanto los seres humanos como el resto de los animales es aplicar el *razonamiento por analogía*, y sólo cuando éste falla replanteamos el problema mediante un análisis de medios-fines. Así, si el mono tiene experiencia propia u observada en otro primate de saltar para alcanzar los plátanos, lo más probable es que su primer intento sea copiar esa estrategia. Sólo tras haberse dado cuenta de que en esa ocasión particular están demasiado altos como para llegar con un salto se planteará buscar medios diferentes como los de la silla y el palo. El cerebro, más que una máquina de razonamiento puro, es un comparador de patrones. Newell y Simon estaban al tanto de este hecho, pero en los años 50 y principios de los 60 los ordenadores no tenían la potencia suficiente para almacenar y manipular amplias bases de datos contenedoras de experiencias previas, y esa limitación tecnológica sólo les permitía explorar estrategias como la de medios-fines (Crevier, 1993, p. 147). Las estrategias que, como la de medios-fines, permiten resolver problemas de cualquier tipo sin requerir conocimiento específico del problema en cuestión se denominan *métodos débiles* (*weak methods*), frente a los *métodos fuertes* (*strong methods*), que sí dependen de una base de datos. Tengamos presente que se trata de una distinción que no tiene nada que ver, y por tanto no hay que confundir, con la de IA fuerte e IA débil (Coppin, 2004, p. 5).

En 1967, una década después de su primera ejecución, el GPS fue abandonado debido, según el propio Newell, a que el programa adolecía de un diseño que sólo lo hacía apto «para resolver problemas modestos cuya representación no fuera

demasiado elaborada» (Dreyfus, 1992, p. 96). A pesar de la decepción del GPS, y de todas los que vendrían después, el Carnegie Tech, bajo la influencia de Newell y Simon, se convirtió en el principal centro de investigación de IA humana, desarrollando proyectos como el SAD SAM de Robert Lindsay, un programa de comprensión del lenguaje natural basado en la aplicación de scripts que podría considerarse un antecesor del SAM de Roger Schank. Sin embargo, la mayoría de los otros asistentes a Dartmouth se decantó por el enfoque diametralmente opuesto, aunque igualmente fracasado (Ibíd., p. 149), de la IA ajena.

En esta otra línea, en 1957, en el Lincoln Laboratory del MIT, Marvin Minsky se unió a Oliver Selfridge para ayudarle a dar los últimos retoques a *Pandemonium*, el programa que definió los principios de diseño que una década después serían copiados por Robert Lindsay y Edward Feigenbaum del Carnegie Tech para crear el primer sistema experto. Mientras que los programas informáticos suelen consistir en secuencias de instrucciones que se ejecutan en un orden determinado, Pandemonium estaba compuesto por una multiplicidad de funciones, bautizadas por Selfridge como "demonios", que eran independientes entre sí. Dada una información, ésta era expuesta en una memoria común a la que todos los demonios tenían acceso para que cada uno de ellos la examinase en busca de determinados rasgos. Cada demonio "gritaba" con mayor o menor intensidad en función de lo seguro que estaba de haber reconocido el rasgo que le era familiar, y un gran demonio maestro elegía el grito más fuerte (Franklin, 1995, p. 234). Por ejemplo, aplicada esta estrategia al reconocimiento de textos, se programarían casi una treintena de demonios, cada uno especializado en reconocer una letra del abecedario. Si la letra en cuestión fuera una F, gritarían los demonios especializados en reconocer letras gráficamente similares, tales como la E, la F, la P e incluso la R, pero el grito más fuerte de todos sería el de la F, por lo que el demonio maestro resolvería que el símbolo es una F.

Mientras en el Lincoln Laboratory se desarrollaba Pandemonium, cerca de allí, en el Departamento de Ingeniería Eléctrica del MIT, John McCarthy dirigía dos proyectos cruciales para impulsar el avance de la IA y de la informática en general: el

LISP y el tiempo compartido. *LISP*, abreviatura de *LISt Processing* (procesamiento de listas), es un lenguaje de programación de alto nivel basado en el cálculo lambda con el que Alonzo Church demostró lo mismo que Turing con su máquina: que el problema de la decisión es irresoluble. Desde su publicación en 1958 hasta mediados de los 80, LISP fue el lenguaje preferido por la mayoría de los investigadores de la IA. En cuanto al *tiempo compartido*, McCarthy diseñó un sistema que permitía a varios usuarios utilizar simultáneamente la misma computadora desde terminales remotos parecidos a máquinas de escribir. Dado que un usuario frente a un terminal no está demandando ciclos de computación constantemente, esos descansos eran aprovechados por el sistema para realizar las tareas solicitadas desde otros terminales. Siendo los ordenadores de la época artefactos escasos por su alto precio, el invento de McCarthy permitió multiplicar las horas frente al ordenador a las que tenía derecho cada profesor y cada alumno del MIT. Su sistema de tiempo compartido no tardó en ser imitado por otras universidades y fabricantes como IBM.

En 1958, una vez terminado Pandemonium, Minsky abandonó el Lincoln Lab y se unió a McCarthy para fundar el *Grupo de Inteligencia Artificial del MIT*, germen de lo que diez años más tarde, en 1968, se convertiría en el *Laboratorio de IA del MIT*, y que en la actualidad forma parte del CSAIL (*Computer Science and Artificial Intelligence Laboratory*). Sin embargo, Minsky y McCarthy pronto se distanciaron por una diferencia de enfoque irreconciliable, y en 1962 McCarthy se trasladó a la Universidad de Stanford para dirigir allí su propio laboratorio de IA. Mientras que McCarthy era partidario de la rama logicista, Minsky defendía la antilogicista (Crevier, 1993, p. 64). La diferencia entre ambas consiste en una diferencia de perspectiva similar a la que separa a los algoritmos de la heurística. Todo lo que hace una computadora electrónica es manipular proposiciones formales de acuerdo a un conjunto de reglas lógicas. La rama *logicista* sostiene que esas proposiciones y sus transformaciones han de reflejar la estructura y el funcionamiento de la mente, mientras que la *antilogicista* defiende que no es necesario que representen nada real dado que son sólo partes de otros procesos de mayor nivel que sí son los que podrían reflejar la mente.

Con la apertura del laboratorio de Stanford se completó el triángulo de los grandes centros de investigación de IA en los Estados Unidos: el del MIT, dirigido por Minsky y Papert, con un enfoque de IA ajena y antilogicista; el de Carnegie Tech, dirigido por Newell y Simon, partidarios de la IA humana; y el de Stanford, dirigido por McCarthy, defensor de la IA logicista.

Los militares enseguida comenzaron a invertir en la emergente disciplina de la IA. En plena Guerra Fría, el Pentágono estaba muy interesado en la posibilidad de construir máquinas inteligentes capaces de realizar automáticamente y en poco tiempo tareas como, por ejemplo, transcribir conversaciones telefónicas en ruso y traducirlas al inglés (Ibíd., p. 110). Para propósitos como éste, en el verano de 1963 el MIT recibió algo más de 2 millones de dólares, de los que un tercio fue destinado al Grupo de IA. La entidad gubernamental emisora de los cheques y supervisora de su utilización era la recién creada *ARPA*, acrónimo de *Advanced Research Projects Agency* (Agencia de Proyectos de Investigación Avanzados), que en 1972 cambiaría su nombre por el de *DARPA*, añadiendo delante el eufemismo *Defence* (Defensa). El proyecto global de ARPA era conocido por su nombre en clave *MAC*, siglas tanto de *Machine-Aided Cognition* (Cognición Auxiliada por la Máquina) como de *Multiple Access Computer* (Computadora de Acceso Múltiple). MAC proporcionó al MIT 3 millones de dólares anuales durante mucho tiempo, dinero del que Minsky invirtió su parte en contratar a los jóvenes más sobresalientes para formarlos en colaboración con el nuevo codirector del Grupo de IA, Seymour Papert, que llegó en 1963 para ocupar el puesto vacante dejado por McCarthy.

Micromundos

A principios de los 60 los investigadores se dieron cuenta de que, si querían avanzar en la creación de máquinas con capacidad de aprendizaje, había una cuestión más fundamental que debían abordar primero: la representación del conocimiento, pues poco podría aprender un programa que, como el GPS, sólo admitiera información

codificada en su propio lenguaje por un operario humano. Para indagar en la representación del conocimiento, dadas las exiguas prestaciones de las computadoras electrónicas de la época (Ibíd., p. 147), los investigadores se vieron obligados a adoptar la estrategia de los *micromundos*, consistente en considerar sólo pequeñas parcelas de la realidad, con poca variedad de entidades y operaciones posibles.

En el ámbito de la representación del lenguaje, destacaron programas como ANALOGY de Tom Evans, STUDENT de Daniel Bobrow y SIR (*Semantic Information Retrieval*) de Bertram Raphael, todos ellos estudiantes de Marvin Minsky. Pero el más impresionante de todos, por su alto grado de eficacia, fue el de las *redes semánticas* de Ross Quillian, un alumno de doctorado de Herbert Simon en Carnegie. Mientras que la mayoría se esforzaba por crear inteligencias artificiales de comprensión del lenguaje natural adoptando el influyente enfoque de Noam Chomsky basado en el análisis exclusivo de la sintaxis, Quillian decidió abrir sus investigaciones a la semántica (Jackson, 1986, p. 54). Para decidir el significado de términos ambiguos, Quillian desarrolló la idea de los nodos de intersección. Supongamos la sentencia siguiente: "Estuve esperándote en el banco". ¿Se refiere a un banco de sentarse o a la sucursal de una entidad financiera? En las redes de Quillian "esperar" activaría las palabras contenidas en su definición, y éstas a su vez activarían las palabras contenidas en sus respectivas definiciones, y así sucesivamente. Lo mismo sucedería con el término "banco" en sus dos acepciones de "banco de sentarse" y "banco de dinero". Cuando una palabra es activada simultáneamente por "esperar" y por alguna de las acepciones de "banco", se dice que ella es un *nodo de intersección*, y se mide la distancia que separa a ambos términos contando el número de saltos que ha habido que dar en la cascada de definiciones. La probabilidad del significado de una palabra es mayor cuantos más nodos de intersección active en convergencia con los significados de las demás y cuanto menor sea la distancia entre los nodos y las palabras.

Un mecanismo muy ingenioso, sin duda, pero las redes semánticas de Quillian tenían muchos defectos. Desde el punto de vista técnico, el programa consumía tantos recursos que sólo dejaba espacio para 20 definiciones en la computadora en la que se

probó en su día. En lo epistemológico, las definiciones consistían en largas listas de enunciados que pretendían, sin éxito, agotar todos los usos posibles del término en cuestión (Ibíd., p. 113), y además el procedimiento de determinación del significado ignoraba el contexto. Esto último es importante, porque el significado más probable de "banco" en la oración anterior no es el mismo cuando es pronunciada en el curso de una conversación entre dos economistas, que pueden estar refiriéndose al Banco de España, que cuando es pronunciada por dos ancianos jubilados. A pesar de todo, las redes semánticas de Quillian fueron una gran innovación en su momento.

En el ámbito de la representación del mundo físico, la conversión de imágenes reales captadas por una cámara a representaciones simbólicas manipulables por una computadora se reveló como una tarea tan extremadamente difícil que, tras los primeros intentos fallidos llevados a cabo por Gerald Sussman por encargo de Minsky, éste y Papert decidieron restringirla al reconocimiento de formas geométricas simples. Dentro de esa estrategia de micromundos denominada *Blocks Micro World* (Micro Mundo de Bloques) cabe mencionar las aportaciones de estudiantes del MIT como Larry Roberts, Adolfo Guzman y Patrick Winston. Éste último, autor por cierto del libro de texto clásico del que antes tomamos la definición del algoritmo Minimax, demostró tal brillantez con su programa de reconocimiento de formas geométricas inspirado en la técnica de las redes semánticas de Quillian, que fue nombrado director del Laboratorio de IA por los propios Minsky y Papert.

La última gran IA de micromundos de bloques fue SHRDLU de Terry Winograd, un estudiante de Papert. Finalizada en 1969, SHRDLU, que debe su impronunciable nombre a las letras de la segunda columna de una máquina de linotipia, reunía habilidades tanto lingüísticas como de representación del mundo físico, aunque de un mundo físico virtual. Consistía en un brazo robótico simulado que manipulaba un conjunto de poliedros sencillos a petición del usuario mediante el lenguaje natural. Así, por ejemplo, se le podía pedir que moviese la pirámide roja para situarla encima del cubo azul. Si el cubo azul tenía encima una pirámide amarilla, entonces SHRDLU comprendía que la orden era imposible y lo comunicaba. Sus técnicas de

reconocimiento de patrones y de palabras clave eran superiores a todo lo visto hasta el momento, gracias a lo cual se convirtió en la primera computadora capaz de intercambiar información significativa en lenguaje natural (Crevier, 1993, p. 99). Para programar SHRDLU Winograd utilizó PLANNER, un lenguaje basado en LISP que permitía trazar planes, fijar metas y realizar aserciones de una manera relativamente sencilla. Los defensores de la IA humana, encabezados por Newell y Simon, se apresuraron a arremeter contra SHRDLU argumentando que PLANNER carecía de una estructura que ayudase a profundizar en la comprensión de las operaciones mentales humanas que subyacen al lenguaje.

Entretanto, en Stanford estaban ocupados gastando el dinero de DARPA en un proyecto no menos útil para fines militares y mucho más ambicioso: un robot llamado *Shakey* capaz de desplazarse por el mundo físico real. Ensamblado en 1969, Shakey contaba con ruedas y varios dispositivos receptores de información, tales como una cámara, un medidor de distancias y un detector de baches. Era parecido a los robots Andros Mark V empleados por algunos cuerpos de policía para desactivar bombas, sólo que sin brazo mecánico y con un cierto nivel de autonomía en vez de ser controlado a distancia, como corresponde a un sujeto inteligente. Este enfoque de la IA, consistente en construir máquinas dotadas de movilidad y órganos sensoriales para que exploren el mundo de manera similar a como lo hacemos los seres humanos, se denomina *IA situada*, frente a la *IA abstracta*, que aspira a crear máquinas pensantes incorpóreas en tanto que confinadas en una computadora inmóvil y con entradas de información seleccionadas (Copeland, 2004, p. 439; Haugeland, 1996, p. 25).

En cualquier caso, situada o abstracta, la IA no consiguió en los años 60 representar el conocimiento tal y como lo hacemos los seres humanos. La razón es que todas las técnicas empleadas eran *ad hoc* (Dreyfus, 1992, p. 1), es decir, diseñadas específicamente para que la máquina en cuestión pudiera habérselas con ciertos problemas predeterminados de un dominio restringido, ya fueran bloques de colores o formas geométricas planas, pero no para representar cualquier tipo de fenómeno, una flexibilidad que es justo la característica distintiva de la cognición humana. Las

soflamas de Minsky elogiando a SHRDLU como un gran avance hacia la representación de cualquier tipo de conocimiento incurrían en la denominada *falacia del primer paso exitoso* (*fallacy of the successful first step*). Esta falacia, según el filósofo judío de origen polaco Yehoshua Bar-Hillel, consiste en reclamar que el hecho de haber dado un primer paso exitoso es razón suficiente para creer que otros pasos similares conducirán a la meta deseada (Ibíd., p. 147). El elogio de Minsky a SHRDLU es, dice Hubert Dreyfus, como si alguien se subiera a un árbol y proclamase haber realizado un progreso significativo para alcanzar la Luna (Ibíd., p. 100).

Redes de neuronas

Al suscribir los participantes de Dartmouth la HSS y al ser ellos los que dirigieron el curso de la nueva disciplina durante las primeras décadas, la IA se limitó en sus inicios a la IA simbólica, que recordemos que es el programa de investigación que aspira a crear una IA programando una máquina, típicamente una computadora electrónica, para que manipule símbolos de manera similar a como se supone que lo hace la mente según el paradigma cognitivista. Pero la IA subsimbólica, con su pretensión de replicar o simular el funcionamiento de las redes de neuronas del cerebro, siempre estuvo presente, aunque fuera en un segundo plano. Así como el supuesto nuclear de la IA simbólica es la HSS, el de la IA subsimbólica es lo que podríamos denominar como la *hipótesis de la semejanza aproximada* (*rough resemblance hypothesis*), o HSA, según la cual para producir inteligencia mediante redes de neuronas artificiales, éstas no tienen por qué ser idénticas a las naturales, sino sólo guardar un cierto (*rough*) parecido (Franklin, 1995, p. 122).

En el capítulo cuarto hablamos ya sobre el modelo de la neurona propuesto por el neurólogo Warren McCulloch y el lógico Walter Pitts en 1943, según el cual las operaciones de una célula nerviosa y sus conexiones con otras formando redes de neuronas pueden ser representadas mediante un modelo lógico (Gardner, 1985, p. 34). Las neuronas se activan y a su vez activan a otras neuronas del mismo modo en

que las proposiciones lógicas pueden implicar otras proposiciones. En palabras de sus autores, lo que pretendieron hacer fue interpretar el cerebro como una gran máquina de Turing (Copeland, 2004, p. 408), una afirmación que contribuyó a la consolidación de la metáfora computacional (Crevier, 1993, p. 30). Pues bien, en 1958 el psicólogo norteamericano Frank Rosenblatt, antiguo compañero de pupitre de Marvin Minsky en los años 40 en una escuela del Bronx, presentó el *perceptrón*, una red de neuronas que combinada el modelo de McCulloch y Pitts con el Pandemonium de Oliver Selfridge (Ibíd., p. 103). El perceptrón consistía en una capa de las neuronas de McCulloch y Pitts flanqueada por encima y por debajo por dos capas adicionales. Las neuronas de la primera capa eran sensores equipados con células fotosensibles que hacían las veces de los demonios de Selfridge, enviando un grito en forma de señal eléctrica proporcional a la intensidad del estímulo lumínico captado. Las neuronas de la segunda capa recibían dichas señales y operaban con ellas realizando una *suma ponderada* (*weighted sum*). Si el resultado excedía una determinada magnitud denominada *umbral* (*threshold*), entonces enviaban una señal a las neuronas de la tercera y última capa, la de salida, que habitualmente convertía la señal eléctrica en un resultado simbólico inteligible para el operario, como números o letras.

Una década antes, en 1948, Turing había diseñado ya una red de neuronas pero, a diferencia de Rosenblatt, no parece probable que se hubiera inspirado en el trabajo previo de McCulloch, a tenor de la mala opinión que el genio inglés tenía de él (Copeland, 2004, p. 408), y lo que es más importante, en el modelo de Turing las conexiones entre neuronas no estaban reguladas por sumas ponderadas, sino por modificadores de dos posiciones: intercambio e interrupción (Turing, 1948, p. 418). La idea de utilizar sumas ponderadas como mecanismo regulador de las conexiones entre neuronas fue de Wesley Clark y Belmont Farley del MIT, quienes en 1954 la incluyeron como característica de la primera simulación informática de una red de neuronas jamás realizada (Copeland, 2004, p. 406). Rosenblatt implementó en su perceptrón la ocurrencia de Clark y Farley para reflejar la diferencia de peso (*weight*), fuerza o conductividad que puede haber entre dos neuronas.

En una suma ponderada la señal entrante que va desde la neurona presináptica A_1 hasta la postsináptica B es multiplicada por un *peso* (*weight*) que refleja la fuerza de la conexión sináptica entre ambas. Gracias a ello es posible que A_1 por sí sola produzca una excitación de B suficiente para que ésta sobrepase el umbral y dispare hacia la tercera capa aun a pesar de que su señal sea débil, mientras que una neurona A_2 que también proyecta sobre B pero a través de una conexión sináptica más débil necesitaría de una señal más fuerte para conseguir el mismo efecto, o incluso puede que por sí sola A_2 nunca sea capaz de excitar a B para que alcance el umbral. La diferencia de fuerza entre las conexiones sinápticas es una característica fundamental del sistema nervioso, y está recogida por el principio de Hebb, el cual establece, recordemos, que cuando una neurona A participa repetida o persistentemente en la excitación o inhibición de otra neurona B, entonces acontece algún tipo de proceso o cambio metabólico en una o ambas que incrementa la eficacia de A para excitar o inhibir a B (Kandel, Schwartz & Jessell, 1995, p. 681). Uno de los méritos de Rosenblatt fue la invención de un mecanismo para ajustar los valores de los pesos, es decir, para ajustar la eficacia con la que una neurona excita o inhibe a otra.

Con el paso del tiempo se demostraría que las redes de neuronas con pesos bien ajustados son más eficaces que las inteligencias artificiales simbólicas en el reconocimiento de ciertos tipos de patrones. Pero, para desgracia de Rosenblatt, su perceptrón fue socialmente aplastado por Minsky y Papert con la publicación en 1969 del libro *Perceptrons: An introduction to computational geometry*. Es paradójico que fuera Minsky el principal instigador de los ataques contra el perceptrón, y no sólo porque fue amigo de Rosenblatt cuando eran jóvenes, sino porque el propio Minsky inició su andadura en la IA desde el enfoque subsimbólico. En el verano de 1951, cinco años antes de la conferencia de Dartmouth, Minsky construyó con la ayuda su compañero de universidad Dean Edmond una red de neuronas. El dinero, unos 2.000 dólares, lo consiguieron gracias al visto bueno de George Miller, uno de los fundadores del paradigma cognitivista, al que Minsky conoció en su paso por Harvard. Construida con 300 tubos de vacío y el piloto automático de un bombardero B-24, la red contaba

con 40 neuronas artificiales (Crevier, 1993, p. 35). Para decepción de Minsky, aquel rudimentario artefacto no era capaz de razonar ni de trazar planes, pues para realizar operaciones intelectuales tan complejas, pensó él, haría falta una red enorme que no estaba a su alcance. Por eso abandonó el enfoque de la IA subsimbólica y, poco después, cuando aparecieron las primeras computadoras electrónicas comerciales como la 701 de IBM, se pasó al de la IA simbólica.

Ciertamente, el perceptrón padecía varios defectos, siendo el principal que las tres capas eran en realidad dos, pues las neuronas de la segunda proyectaban directamente sobre las de la tercera de manera directa, en relación de una a una y sin intermediación de pesos. Al ser sólo dos las capas efectivas, esto daba lugar a que el perceptrón no pudiese ejecutar la operación lógica XOR, o disyunción exclusiva, que es aquella en la cual el resultado es verdadero sólo cuando los dos términos tienen distinto valor de verdad. En vez de adoptar una actitud constructiva para superar éste y otros defectos, como harían David Rumelhart y James McClelland años después, Minsky y Papert decidieron hundir a Rosenblatt acusándole de cargos tan graves como la falta de cientificidad, y con él hundieron durante más de una década a la IA subsimbólica, pues Rosenblatt murió poco después, y hasta los años 80 nadie volvió a desviarse públicamente de la IA simbólica para investigar las redes de neuronas.

Primer invierno

A principios de la década de los 70 las computadoras electrónicas aumentaron considerablemente sus prestaciones gracias al salto de la segunda generación de estas máquinas, basada en transistores, a la tercera, caracterizada por los circuitos integrados o microchips, mucho más rápidos, pequeños y fiables. Lejos de propiciar un avance en la IA, la tercera generación fue un contratiempo para los investigadores, pues de repente se quedaron sin la excusa de la falta de potencia de los ordenadores que hasta el momento habían venido esgrimiendo para encerrarse en los micromundos en vez de abordar el problema de la representación del mundo real en

toda su amplitud (Ibíd., p. 147). Edward Feigenbaum, de quien hablaremos más adelante por ser el creador del primer sistema experto, acusó así a sus antiguos maestros: «Ustedes trabajan en problemas pueriles. El ajedrez y la lógica no son más que juguetes. Si los resuelven, sólo habrán resuelto puerilidades, eso es todo lo que habrán hecho. Vayan al mundo real y resuelvan los problemas del mundo real» (Gardner, 1985, p. 181). El verdadero obstáculo para crear máquinas capaces de habérselas con el mundo real no era la falta de potencia de los ordenadores, sino una serie de problemas epistemológicos que, hasta ese momento, habían sido ignorados porque la IA, bajo la influencia del paradigma cognitivista, se había conformado con replicar la inteligencia en condiciones ideales de laboratorio, como las de los micromundos, purgadas de factores ambientales complejos. Los primeros en señalar los problemas epistemológicos que surgen cuando una IA se enfrenta al mundo real fueron John McCarthy y el matemático Patrick Hayes.

En 1969 McCarthy y Hayes publicaron *Some philosophical problems from the standpoint of artificial intelligence*, un artículo en el que advertían de que la IA no avanzaría en el problema de la representación del conocimiento hasta que cobrase conciencia de la necesidad de afrontar una serie de cuestiones epistemológicas extremadamente difíciles que ni los más grandes filósofos habían conseguido resolver en veinticinco siglos. De entre todas esas cuestiones, McCarthy y Hayes destacaban dos: el problema del marco (*frame problem*) y el que más tarde daría en llamarse el problema de la cualificación (*qualification problem*). Del primero ya hablamos en el capítulo anterior. Su nombre se debe a que cualquier cambio ha de ser relativizado respecto a un marco de referencia, y se trata del problema de cómo actualizar una gran base de datos interconectados entre sí en la cual la variación de uno solo de ellos puede implicar la necesidad de modificar muchos otros (Copeland, 1993, p. 143). Por poner un ejemplo sencillo, si un sistema de seguridad inteligente monitorizase a un hombre que lleva una llave en la mano, cuando el hombre cambiara de habitación el sistema debería ser capaz de inferir que la llave también ha cambiado de habitación y actualizar la base de datos en consecuencia (Franklin, 1995, p. 116).

En cuanto al *problema de la cualificación*, está relacionado con el anterior, y se refiere a la imposibilidad de detallar en un listado todas las condiciones que deben cumplirse para que una regla general sea válida (McCarthy & Hayes, 1969, p. 34). Recordemos la historia de Eugene Charniak que vimos en el capítulo anterior: «Jane fue invitada a la fiesta de cumpleaños de Jack. Ella se preguntó si le gustaría una cometa. Fue a su habitación y agitó su cerdito, pero no sonó» (Minsky, 1975, p. 103). Jane soluciona el problema del dinero y el texto continúa así: «Penny y Jane fueron a la tienda a comprar regalos. Jane decidió comprar una cometa. "No lo hagas", dijo Penny. "Jack ya tiene una cometa, y te hará que la devuelvas"» (Dreyfus, 1992, p. 57). Para que una IA comprendiese la última frase debería conocer la regla general de que, normalmente, si uno tiene un objeto, no quiere tener otro igual. Ahora bien, excepciones a esta regla serían las galletas, las canicas, las piedras preciosas, y así hasta completar una lista inimaginablemente extensa. A su vez, hay excepciones a las excepciones, como que mil canicas son suficientes para un niño. Y también hay excepciones a las excepciones de las excepciones, como el caso del niño que tiene mil canicas porque su padre lo ha introducido en el coleccionismo de estos objetos.

La cuestión es que los seres humanos no tenemos explícitamente en la memoria listados de este tipo, y sin embargo sabemos cuándo funciona y cuándo no la regla general de que la gente no quiere tener dos objetos iguales. En cambio, una IA simbólica sí necesita un listado, pero es imposible proporcionárselo porque el programador no puede prever todas las situaciones en las que una regla general dejará de funcionar. Los listados de excepciones desembocan en un *regreso infinito* de excepciones a las excepciones, como se observa en el ejemplo de las canicas. Ambos problemas, el del marco y el de la cualificación, continúan siendo, a día de hoy, dos de los principales obstáculos para la construcción de inteligencias artificiales simbólicas.

Las críticas de McCarthy y Hayes supusieron un duro revés para los investigadores de la IA. No obstante, fueron asimiladas, dado que provenían de una autoridad en la materia como era McCarthy. Las que no fueron asimiladas, sino rechazadas con un odio africano, fueron las críticas formuladas por el filósofo de

orientación fenomenológica Hubert Dreyfus. Los argumentos de Dreyfus, dirigidos a refutar la posibilidad técnica por principio de la IA simbólica, los examinaremos en el próximo capítulo. Baste por ahora con apuntar que las represalias que él sufrió en el MIT fueron terribles. En palabras suyas: «El rechazo era tan absoluto que los estudiantes y los profesores que trabajaban en proyectos de robótica no se atrevían a sentarse conmigo en el comedor a la hora del almuerzo por temor a meterse en problemas con alguno de sus superiores. Cuando Joseph Weizenbaum, el único profesor que tenía dudas (sobre la IA), quería discutir algún asunto conmigo, me citaba para que nos viésemos en su casa a las afueras» (Crevier, 1993, p. 122).

Las voces de McCarthy, Hayes, Dreyfus y Weizenbaum contra la IA por un lado, y por el otro la falta de progreso efectivo de los investigadores en la construcción de máquinas verdaderamente inteligentes que tuviesen alguna utilidad en el mundo real fueron las causas por las cuales en 1974 DARPA decidió reducir drásticamente la financiación. Al otro lado del océano, las autoridades británicas hicieron lo mismo, y así fue cómo comenzó en esa fecha el *primer invierno de la IA* (Kaku, 2011, p. 108). Irónicamente, el autor del informe que advirtió al gobierno británico de la inviabilidad de la IA a corto plazo fue James Lighthill, un profesor lucasiano, como también había sido otro profesor lucasiano, George Biddell Airy, quien en el siglo XIX recomendó a la reina Victoria cortarle la financiación a Charles Babbage por considerar que su máquina de diferencias nunca sería realmente útil y eficaz. Sin dinero público a ambos lados del Atlántico, la IA permaneció bajo mínimos hasta que en la década de los 80 inventó un producto con aplicaciones comerciales lucrativas: los sistemas expertos.

Sistemas expertos

Edward Feigenbaum, antiguo alumno de Herbert Simon en Carnegie, inició en 1965 en la Universidad de Stanford un proyecto llamado *DENDRAL* en colaboración con Robert Lindsay y el premio Nobel de Medicina Joshua Lederberg (Jackson, 1986, p. 19). Tras más de diez años de desarrollo, DENDRAL se convirtió en el primer sistema

experto de la Historia. Su utilidad era averiguar la estructura de moléculas orgánicas complejas. El diseño de DENDRAL, basado en la arquitectura no jerarquizada del Pandemonium de Oliver Selfridge, consistía en un conjunto de oraciones condicionales del tipo "Si A entonces B" (A→B), de tal manera que, si una molécula cumplía varias de ellas, se infería cuál debía de ser su estructura. Durante el desarrollo, Feigenbaum se dio cuenta de que DENDRAL padecía un error de diseño irreparable, y es que las oraciones condicionales y las reglas que decidían su aplicación estaban unidas en un solo programa, escrito por cierto en LISP. En colaboración con Bruce Buchanan, otro antiguo alumno de Carnegie que también se había trasladado a Stanford, Feigenbaum encontró la solución en unos trabajos anteriores de Simon y Newell acerca de un modelo alternativo de la cognición humana: los sistemas de producción.

Los *sistemas de producción* (*production systems*) fueron inventados en 1943 por el matemático estadounidense de origen polaco Emil Post (Franklin, 1995, p. 72). Se trata de sistemas formales equivalentes a la máquina de Turing, es decir, que cualquier tarea que pueda ser realizada por una máquina de Turing puede ser realizada por un sistema de producción, y viceversa, gracias a lo cual pueden ser ejecutados en computadoras electrónicas de propósito general (Ibíd., p. 79). Un sistema de producción se compone, al igual que DENDRAL, de un conjunto de oraciones condicionales denominadas *reglas de producción* (*production rules* o, simplemente, *productions*) y otro conjunto de instrucciones llamada *estructura de control* (*control structure*) o *motor inferencial* (*inference engine*) que decide cuál es la regla que, partiendo de un estado de cosas inicial, debe ser aplicada en cada momento para alcanzar la meta o estado de cosas final. La importante diferencia entre DENDRAL y un sistema de producción es que en un sistema de producción la estructura de control es independiente y está separada de las reglas de producción. Además, en un sistema de producción se distingue un tercer elemento (Jackson, 1986, p. 31): la *base de datos global* (*global database*), también llamada *memoria de trabajo* (*working memory*), que, en términos del análisis de medios-fines, se define como el lugar donde se representa el estado de cosas actual y las diferencias que lo separan de la meta.

Conviene señalar que, en rigor, no todos los sistemas expertos son sistemas de producción, pues los formalismos contenedores de la información no tienen por qué ser necesariamente reglas de producción (Ibíd., p. 30), sino que también pueden ser, entre otros, la lógica de predicados y los objetos estructurados, los cuales fueron la elección de Quillian para sus redes semánicas (Ibíd., p. 53). Sin embargo, aquí sólo nos vamos a referir a los sistemas expertos basados en sistemas de producción, por ser los más extendidos y por su importancia histórica dado que fueron los primeros.

El primero de todos, presentado en 1972, fue *MYCIN*. Obra de Edward Shortliffe, un estudiante de Buchanan en Stanford, MYCIN diagnosticaba infecciones de la sangre. Una de sus características más reseñables era que, al igual que muchos médicos humanos, para realizar sus inferencias diagnósticas empleaba la técnica del *razonamiento regresivo* (*backward chaining*). Dado un conjunto de reglas de producción, éste puede ser siempre dispuesto en forma de árbol de decisión (Crevier, 1993, p. 154). Una vez MYCIN había recorrido el árbol por un determinado camino para llegar desde los síntomas hasta el diagnóstico, aplicaba el razonamiento regresivo, del diagnóstico a los síntomas, para asegurarse de que todas las inferencias eran correctas. Otra virtud destacada de MYCIN era que sus inferencias podían ser estadísticas. Puesto que la medicina no es una ciencia exacta, a veces los médicos tienen que trabajar con probabilidades para explicar las causas de un síntoma. Posteriormente, la estructura de control de MYCIN fue aislada y ofrecida con el nombre de EMYCIN (del inglés *Empty* MYCIN, MYCIN vacío) a otros investigadores para que desarrollaran sistemas expertos simplemente uniéndola a nuevos conjuntos de reglas de producción (Jackson, 1986, p. 107).

Los primeros sistemas expertos, como DENDRAL y MYCIN, marcaron un hito en el devenir de la IA, porque eran máquinas de métodos fuertes, es decir, con una inteligencia basada en el conocimiento (Crevier, 1993, p. 156), a diferencia de los programas anteriores basados en métodos débiles que, como el GPS de Newell y Simon, no tenían ningún conocimiento de nada y confiaban la solución de cualquier tipo de problema a la recta aplicación de la razón por sí sola, tal y como le habría

gustado a Descartes. Por estar basados en su mayoría en la arquitectura de sistemas de producción, los sistemas expertos tienen varias limitaciones y virtudes, de entre las que, por el momento, destacamos sólo una de cada. La limitación más ostensible es que permanecen confinados en micromundos, dado que sólo pueden resolver problemas relativos al dominio codificado en sus reglas de producción. Por eso, como dijimos en el primer capítulo, son el equivalente en el reino de las máquinas a los *deficientes geniales* o *síndrome de savant* en el de los seres humanos: personas incapaces de realizar tareas tan sencillas como atarse los cordones o hablar de lo que cenaron ayer pero que, en cambio, tienen una habilidad extraordinaria para las matemáticas, la pintura, la música o el ajedrez. La virtud es que, al ser sus reglas de producción no jerarquizadas a semejanza de los demonios de Selfridge, en cualquier momento pueden añadirse otras adicionales para ampliar los conocimientos del sistema sin necesidad de alterar las preexistentes.

Para determinar cuáles debían ser las reglas de producción de un sistema experto, enseguida apareció una nueva disciplina: la *ingeniería del conocimiento* (*knowledge engineering*). Su objetivo es extraer el conocimiento que hay en la mente de un experto humano y transferirlo a la memoria del sistema informático. Las dificultades para realizar esa tarea son, por lo menos, dos. La primera es que el ingeniero no suele ser experto en el área del conocimiento en cuestión, como por ejemplo la medicina o la química, y la segunda es que el experto humano entrevistado no posee todo su conocimiento de manera consciente, tal y como apuntamos en el capítulo anterior. Para diagnosticar a un paciente, un médico emplea reglas que es capaz de formular, pero también emplea otras que ni él mismo sabe que las sabe. Muchas son de sentido común, y son las que le hacen darse cuenta, por ejemplo, de que si un informe dice que una paciente de maternidad pesa 25 kilos y tiene 70 años, lo que ha sucedido es que el peso (70) y la edad (25) han sido intercambiados por error. Debido al problema de la cualificación, un sistema experto necesitaría que se le proporcionase un listado de todas las condiciones de validez de cada regla de producción, pero eso es imposible, porque el experto humano no es capaz de

enunciarlas. Sin el conocimiento de esas condiciones, un sistema experto ordenaría que a la paciente de maternidad se le dispensara la medicación adecuada para una mujer de 25 kilos y 70 años. «Conseguir que las máquinas hagan "lo que yo quería decir, no lo que dije" es todavía un problema de primer orden en la investigación de la ciencia computacional» (Ceruzzi, 1998, p. 93).

Con todos sus defectos y virtudes, los sistemas expertos empezaron a ser comercializados a finales de los 70. El primero se conoce con los nombres de *R1* o *XCON*, abreviatura de *eXpert CONfigurer* (*configurador experto*). Diseñado por John McDermott de Carnegie para la Digital Equipment Corporation (DEC), XCON se encargaba de configurar las nuevas computadoras VAX antes de empaquetarlas. La primera versión del programa, ejecutada en 1979, constaba de 300 reglas, y fue creciendo hasta que en la versión de 1984 alcanzó las 3.000 y era capaz de configurar varios modelos de ordenadores (Jackson, 1986, p. 137). XCON ocupa un lugar destacado en la Historia de la IA porque fue la primera vez que esta disciplina salió de los laboratorios y demostró que era capaz de producir algo útil (Crevier, 1993, p. 161). Tras décadas de investigaciones sin arrojar resultados prácticos, la IA empezaba por fin a exhibir rendimiento en términos de racionalidad instrumental.

El iceberg

Cuando se dice vulgarmente que los seres humanos sólo utilizamos el 10% del cerebro, lo que quiere decir esa frase es que sólo utilizamos una pequeña fracción de manera consciente, pero si tenemos en cuenta los procesos inconscientes, en realidad lo utilizamos al 100%. Sin ir más lejos, algo tan racional como leer es un proceso inconsciente en gran medida, pues lo hacemos sin pensar en el significado de cada palabra ni en los efectos modificadores de la sintaxis y la pragmática. Por tanto, al igual que un iceberg, el conocimiento puede dividirse en dos partes: una pequeña que está a la vista y de la que somos conscientes, y otra mucho más grande que está en la base de la anterior y que permanece oculta en el inconsciente. En los años 70, mientras

Feigenbaum, Shortliffe y McDermott desarrollaban los primeros sistemas expertos, otros investigadores como Roger Schank, Marvin Minsky, y David Marr dedicaban sus esfuerzos a elaborar modelos cognitivos que contemplasen la parte oculta del iceberg. Dado que en el capítulo anterior ya nos referimos a la primera teoría de los scripts de Schank, aquí sólo vamos a mencionar las teorías de Minsky y de Marr.

En 1975 Minsky publicó *A framework for representing knowledge*, un artículo en el que proponía que todo acto comprensivo relaciona el objeto de la comprensión con una estructura de memoria denominada *marco* (*frame*). Los marcos son representaciones invariables que contienen las características estereotipadas del objeto en cuestión, ya sea éste un simple dormitorio o algo tan abstracto como una fiesta de cumpleaños. Asociadas al marco hay muchas informaciones que indican cosas tales como la manera de usarlo, qué es lo que va a suceder a continuación o qué hacer en caso de que las expectativas no se cumplan. Un marco, dice Minsky, es como una red de nodos y relaciones: «Los niveles superiores están fijados, y representan cosas que son siempre verdaderas acerca del objeto. Los inferiores tienen varias *terminales* (*terminals*) –ranuras que deben ser rellenadas por instancias específicas o datos. Cada terminal puede especificar condiciones que sus entradas (*assignments*) deben satisfacer. (Las propias entradas son habitualmente submarcos menores). Las condiciones simples son especificadas por *marcadores* (*markers*) que pueden requerir que la entrada de un terminal sea una persona, un objeto de suficiente valor, o un puntero hacia un submarco de cierto tipo» (Minsky, 1975, p. 96).

El artículo concluía con un apéndice en el que Minsky reafirmaba su enfoque antilogicista de la IA argumentando que «la lógica tradicional no puede habérselas con problemas reales y complicados porque no es adecuada para representar aproximaciones a las soluciones –y esto es algo absolutamente crucial» (Ibíd., p. 98). Estas palabras de Minsky resumen una de las conclusiones del capítulo anterior: que el círculo hermenéutico es aproximativo, flexible, impreciso, mientras que el lenguaje formal por el que se rigen las computadoras es extremadamente rígido. Obviamente, el texto de Minsky no gustó a los defensores de la rama logicista, como su antiguo

compañero John McCarthy. Éste le reprochó además la utilización del término *marco* (*frame*) con un significado distinto al que él le había dado previamente en su artículo de 1969 sobre los problemas epistemológicos de la IA. Por su parte, a Herbert Simon le molestó que se le reconociera públicamente a Minsky la invención de los marcos, cuando resulta que, según él, se trataba de un concepto idéntico al de *esquema* (*schema*) ideado mucho antes, en 1956, por él mismo y por Allen Newell (Crevier, 1993, p. 174). Teniendo en cuenta su enfoque de IA humana, es probable que Newell y Simon tomaran prestado ese concepto de Frederick Bartlett, un psicólogo inglés de la primera mitad del siglo XX, quien a su vez debía la noción de esquema a Kant.

Más allá de las críticas particulares, en general a Minsky se le reprochó que su artículo carecía de exactitud. Así como Schank no explica cómo seleccionamos inconscientemente una determinada estructura memorística (Dreyfus, 1992, p. 45), Minsky tampoco detalla cómo se produce la selección del marco más adecuado para cada objeto. En su favor, Minsky argumentó que la pretensión del texto no era agotar exhaustivamente una línea de investigación, sino sólo sugerirla para que otros indagaran en ella más a fondo. De hecho, ésa ha sido la principal contribución de Minsky a la IA durante toda su carrera: impulsar la investigación en direcciones abiertas más que investigar por sí mismo hasta cerrar los temas. En la actualidad, su teoría de los marcos es considerada no como la gran solución a los problemas de la IA simbólica, pero sí como una herramienta útil (Crevier, 1993, p. 175).

En cuanto a David Marr, a mediados de los 70 fue invitado por Minsky y Papert a trasladarse al MIT para continuar desarrollando allí su teoría de la visión. Sus estudios fueron de escasa utilidad práctica, ya que, como neurofisiólogo que era, se dedicó a los niveles más bajos de procesamiento de la información visual, y además no tuvo tiempo para completarlos, pues desgraciadamente murió de leucemia cuando tenía sólo 35 años. No obstante, su obra es considerada de gran influencia (Ibíd., p. 188). Según Marr, la visión es producto de tres representaciones sucesivas (Gardner, 1985, p. 327): bosquejo primario, bosquejo 2½D y representación 3D. El *bosquejo primario* se crea al margen de cualquier conocimiento de alto nivel, y consiste en la

reducción de la imagen a elementos simples, tales como líneas, bordes y manchas. Estos elementos se agrupan en la fase siguiente para dar lugar a la formación de un *bosquejo 2½D* en el que se distingue la forma y el volumen de los objetos. Y, finalmente, los objetos y sus partes componentes son identificados comparándolos con memorias almacenadas para dar lugar a la *representación 3D*, que es la que vemos automáticamente con sólo abrir los ojos.

Como se puede apreciar, Marr describe la visión en términos del mismo proceso bidireccional de abstracción creciente que en el capítulo anterior, a propósito de la teoría de la inteligencia de Jeff Hawkins, señalamos como el algoritmo fundamental no sólo de la visión, sino de toda la corteza cerebral. La corteza, dice Hawkins, es una memoria autoasociativa cuya herramienta computacional común consiste en elaborar representaciones invariables y utilizarlas para realizar predicciones (Hawkins & Blakeslee, 2004, p. 66). La elaboración de representaciones invariables se realiza a partir de las entradas sensoriales que fluyen de abajo a arriba (*bottom-up*), mientras que la predicción acontece por el cruce de dichas entradas con las proyecciones de arriba a abajo (*top-down*). En el esquema de Marr, las dos primeras fases son puramente abstractivas de abajo a arriba, y es en la tercera cuando nuestro conocimiento previo sobre el mundo, almacenado en forma de abstracciones del tipo de los marcos de Minsky, se proyecta de arriba a abajo para determinar por comparación la identidad de lo percibido (Crevier, 1993, p. 188).

Segundo invierno

«Hacía horas que Molly había regresado a la buhardilla; llevaba la estructura del Flatline en el bolso verde, y desde entonces Case había estado bebiendo sin interrupción. Trastornaba pensar en el Flatline como una estructura: una cassette de circuitos ROM que reproducía las habilidades, obsesiones y reflejos de un muerto. [...] –Cómo te va, Dixie. –Estoy muerto, Case. He pasado ya bastante tiempo en este Hosaka como para saberlo. –¿Qué se siente? –No se siente. –¿Te molesta? –Lo que me

molesta es que nada me molesta» (Gibson, 1984, pp. 99 y 132). En estos fragmentos de su novela de ciencia ficción *Neuromante*, William Gibson habla de un soporte informático de memoria, la estructura Flatline, que contiene una copia ejecutable de la mente de un ser humano. Si concedemos que los psiquiatras virtuales como Sigfrid existen desde que Kenneth Colby escribió DOCTOR en los años 60, entonces las personas virtuales como el Flatline existen desde principios de los 80.

El éxito del sistema experto XCON convenció a DEC de la rentabilidad de invertir en ese tipo de programas informáticos. Cuando la tarea de configurar ordenadores se le asignaba a un ser humano, éste resultaba mucho menos productivo que un sistema experto, pues la computadora puede configurar un número indefinido de ordenadores simultáneamente, no cobra un salario, no se queja por trabajar en turnos de 24 horas, no va a la huelga y puedes despedirla cuando quieras sin ningún coste. El resto de empresas no tardaron en seguir los pasos de DEC, y apostaron también por las ventajas de los sistemas expertos. Así, cuando en 1981 la General Electric Company se encontró con el problema de que su mejor ingeniero, David Smith, deseaba jubilarse, la solución adoptada fue aplicarle a Smith las técnicas de ingeniería del conocimiento y codificar sus habilidades intelectuales en un sistema experto (Crevier, 1993, p. 195). La General Electric no sólo consiguió el objetivo prioritario de seguir disponiendo del talento de Smith después de que éste se jubilase, sino que además lo multiplicó en copias que se distribuyeron por las oficinas de la compañía a lo largo de los Estados Unidos. Fue lo mismo que hizo la Campbell's Soup Company con su especialista en hidrostática Aldo Cimino, o la General Motors con su ingeniero Charlie Amble. Copias en discos de la única parte de la mente de los seres humanos que interesa al capital, guiado por la razón instrumental: la productiva.

En los Estados Unidos, siempre a la vanguardia de la tecnología, la mayoría de las grandes corporaciones empresariales puso en marcha divisiones internas de IA con el objetivo de desarrollar sistemas expertos para uso propio. Se calcula que en 1985, entre 150 empresas gastaron un total de 1.000 millones de dólares en dichas divisiones (Ibíd., p. 199). Al año siguiente, 1986, las ventas de hardware y software

relacionado con la IA alcanzaron los 425 millones. Algo más de la mitad de esa cifra correspondía a unas microcomputadoras denominadas *máquinas LISP* (*LISP machines*), diseñadas expresamente para ejecutar a alta velocidad los programas escritos en LISP. El segundo negocio más lucrativo era el de las *herramientas de desarrollo de sistemas expertos* (*expert system development tools*), o simplemente *shells* (conchas, caparazones). Un shell consiste en una estructura de control, una base de datos global y un interfaz de usuario (Ibíd., p. 153), es decir, lo mismo que ofrecía EMYCIN. Para crear sus propios sistemas expertos, las compañías compraban un shell y su división de ingenieros se encargaba de añadirle el conjunto de reglas de producción.

Gracias al éxito de los sistemas expertos a principios de los 80, la IA salió de su primer invierno. La demanda del sector privado aumentaba cada año, y las inversiones públicas a través de DARPA regresaron en forma de proyectos como, por ejemplo, el de un camión militar conducido por un piloto automático. Los ingenieros informáticos estaban tan solicitados, que con frecuencia abandonaban las universidades antes de haber terminado la carrera. Sin embargo, la abundancia duró poco. La IA cayó enseguida en su segundo invierno por culpa de las debilidades inherentes a la arquitectura de los sistemas de producción.

«En 1984 Roger Schank y Marvin Minsky advirtieron de que el entusiasmo por la nueva tecnología estaba fuera de control. Argumentaban que los sistemas expertos se basaban en métodos de programación de veinte años de antigüedad que sólo habían ganado potencia gracias al incremento de la velocidad de los ordenadores» (Ibíd., p. 203). Dicho de otra forma, eran programas que hacían lo mismo que los de veinte años atrás: «tomar decisiones estúpidas, sólo que más rápido» (Ibíd., p. 203). Schank señaló que los sistemas de producción no razonan como lo hacemos los seres humanos (Schank, 1999, p. 230). Nosotros elaboramos reglas generales (*rules of thumb*) abductivamente en base a experiencias particulares, y además somos capaces de razonar sin necesidad de reglas generales, operando directamente sobre las experiencias particulares mediante un proceso denominado *razonamiento basado en casos* (*case based reasoning*), o *CBR*. En cambio, los sistemas de producción no tienen

tales capacidades de aprendizaje y abstracción, sino que sólo razonan en base a reglas generales que además deben serles proporcionadas (Ibíd., p. 215). El problema de esto es que, dado que las reglas de un sistema de producción no están jerarquizadas, la adición de una nueva puede producir efectos imprevistos como resultado de su interacción con las preexistentes.

Es algo análogo a lo que sucede con las leyes del Código Penal o del Código Civil: cuando son pocas, su efecto en forma de sentencias concretas es predecible, pero cuando son muchas dan lugar a que los delincuentes con más dinero para gastar en abogados puedan quedar absueltos. La misión de esos abogados es buscar enrevesadas pero legítimas combinaciones de leyes que arrojen resultados favorables para sus clientes e introducirlas en el conjunto de reglas de producción que el juez utilizará para dictar sentencia. No son pocos los jueces que hacen lo mismo que los sistemas expertos: tomar decisiones estúpidas, sólo que más despacio. Es como si ciertos magistrados también padecieran el problema de la cualificación, pues se comportan ignorando las condiciones de validez de las reglas generales en el mundo real. Volviendo al XCON, cuando en sus inicios constaba de 300 reglas, su comportamiento era bastante fiable y rentable, pero cuando en 1984 alcanzó las 3.000, se hizo necesario tener en nómina a un equipo de 150 personas encargadas de diseñar las nuevas reglas y predecir sus efectos (Crevier, 1993, p. 204). Pagar los salarios de tantos obreros cualificados para mantener en funcionamiento una máquina cada vez menos fiable empezaba a no compensar económicamente.

La solución más inmediata habría sido jerarquizar las reglas de producción, pero eso habría privado a los sistemas de producción de su principal ventaja respecto de los programas convencionales: la modularidad. Así que a mediados de los 80 los ingenieros redoblaron sus esfuerzos es desarrollar técnicas para que las inteligencias artificiales fueran capaces de adquirir conocimiento por sí solas. Se toparon entonces con los problemas señalados por McCarthy y Hayes: el del marco y el de la cualificación, y con muchos otros, como la debilidad de los sistemas expertos para comprender el paso del tiempo y la causalidad. Respecto a la causalidad, Minsky

propuso el *desafío del pato* (*duck challenge*) (Ibíd., p. 206). Supongamos, dice, que un sistema experto aprende que "todos los patos vuelan" y que "Charlie es un pato". Deducirá sin problemas que "Charlie vuela". Pero si a continuación se le informa de que Charlie está muerto, debe ser capaz entonces de *desaprender* que "Charlie vuela", pues los muertos no vuelan. El tipo de lógica requerida para desaprender lo aprendido es la *lógica no monotónica* (*nonmonotonic logic*). Sin embargo, a día de hoy no se logrado todavía que las computadoras se rijan por ella de manera eficaz. El CYC de Douglas Lenat es un ejemplo de IA que funciona, o por lo menos lo intenta, con lógica no monotónica (Copeland, 1993, p. 164).

Hacia finales de los 80 los sistemas expertos entraron en crisis. Los ingenieros no habían sido capaces de enseñarles a aprender por sí solos de manera similar a como lo hacemos los seres humanos, y la acumulación de reglas generales a esas alturas era ya insostenible, tanto por el descenso de la fiabilidad de los sistemas como por el incremento de los costes de personal dedicado a mantenerlos operativos. En consecuencia, fueron muchas las empresas que decidieron prescindir de ellos. El toque de gracia que terminaría por hundir a la IA en su *segundo invierno* a finales de los 80 (Kaku, 2011, p. 109) fue la popularización de las computadoras personales. En 1987 salieron a la venta el Macintosh II de Apple y el 386 para IBM y compatibles, plataformas ambas con una potencia similar a las máquinas LISP pero mucho más baratas. Los desarrolladores de software, incluidos los de IA, decidieron pasarse a ellas utilizando para programarlas el lenguaje *C*, abandonando así el LISP y provocando un desplome de todo el mercado de software y hardware que durante los últimos años había surgido alrededor del lenguaje de programación de John McCarthy.

No obstante, el efecto más importante de los ordenadores personales sobre la IA fue más de tipo psicológico que económico. Las máquinas como el Macintosh no se comercializaban como "cerebros electrónicos" con el propósito de sustituir a la inteligencia humana, sino para servir de apoyo a ésta poniendo a su disposición herramientas como hojas de cálculo, procesadores de texto y programas de diseño gráfico. Esta nueva visión de las computadoras se trasladó a los sistemas expertos. Así,

mientras que en los 80 el objetivo era convertirlos en sustitutos de seres humanos como Smith, Cimino y Amble, en los 90 pasaron a ser simples consultores (Crevier, 1993, p. 213), que es la utilidad que tienen hoy en día. En la terminología de André Robinet, diríamos que se asumió la imperfectibilidad insalvable y se decidió que la condición redoblante del doble era de suficiente provecho.

6.2. La IA subsimbólica en la actualidad

A mediados de los 80, la moderación de las expectativas puestas en los sistemas expertos contribuyó al resurgir de las redes de neuronas. Tras más de una década abandonadas por culpa de las duras críticas de Minsky y Papert contra el perceptrón de Rosenblatt, el programa de investigación de la IA subsimbólica regresó al centro de la escena gracias a dos acontecimientos (Ibíd., p. 215). El primero fue la publicación de un artículo en 1982 en el que el físico John Hopfield describía unas *redes de neuronas autoasociativas* (*autoassociative neural networks*) de cuya interacción emergían habilidades computacionales. El segundo, aún más importante, fue la edición en 1986 de dos volúmenes titulados *Parallel distributed processing* (*Procesamiento distribuido en paralelo*), o simplemente *PDP*, obra de un grupo de investigadores conocidos como *Grupo PDP* bajo la dirección de los psicólogos David Rumelhart de la Universidad de San Diego y James McClelland de Carnegie. El PDP, siglas sinónimas en la actualidad del enfoque subsimbólico o conexionista, consiguió pronto éxitos muy notables que llamaron la atención.

Paul Smolensky, profesor de ciencia cognitiva en la Universidad Johns Hopkins y miembro del Grupo PDP, enumera los siguientes fenómenos cognitivos simulados por las redes de neuronas en sus primeros años (Smolensky, 1989, p. 233): percepción del habla, reconocimiento visual de figuras de papiroflexia, desarrollo de detectores especializados de características, amnesia, generación y análisis gramatical (*parsing*) del lenguaje, afasia, descubrimiento de códigos binarios, programación de redes paralelas masivas, adquisición de la morfofonología del tiempo verbal pasado en inglés

a partir de ejemplos, jugar al tres en raya, inferir el contenido de habitaciones y resolución cualitativa de problemas en circuitos eléctricos simples. Semejantes avances atrajeron a muchos otros investigadores, hasta el punto de que en 1991, según McClelland, en los Estados Unidos había 10.000 personas trabajando en las nuevas redes de neuronas. Desde entonces hasta el presente, la IA subsimbólica ha permanecido activa sin interrupción. Al igual que la IA simbólica, la IA subsimbólica se divide en dos vertientes: la realista de la IA humana y la instrumental de la IA ajena. Veamos, en este orden, el estado en el que se encuentran.

Una de las supercomputadoras más modernas dedicadas a la simulación de redes de neuronas es *Blue Gene*. Su nombre delata su parentesco con Deep Blue, la IA jugadora de ajedrez que derrotó a Kaspárov. El azul (*blue* en inglés) es el color corporativo de IBM, por lo que allá donde se hable de superordenadores con el nombre "Blue", es muy probable que se trate de una máquina de la empresa fundada por Herman Hollerith. Blue Gene cuenta ya con tres generaciones a sus espaldas: Blue Gene/L, presentada en 2004, Blue Gene/P, que data de 2007, y Blue Gene/Q, de 2012. El proyecto que culminaría en el primer Blue Gene comenzó en 1999 con el objetivo de fabricar una gran computadora de procesamiento masivo en paralelo. Dos años después, los militares se dieron cuenta de su importancia estratégica, y el Pentágono se unió como socio a IBM ofreciéndole sus instalaciones del Laboratorio Nacional Lawrence Livermore (LLNL) en California, el centro de investigación de armas secretas más importante de los Estados Unidos (Kaku, 2011, p. 137). En cualquiera de sus modelos, Blue Gene tiene muchas aplicaciones, desde civiles a militares, pero la que más fama le ha dado es quizás la simulación de redes de neuronas.

En 2007 un equipo de científicos dirigido por Rajagopal Ananthanarayanan utilizó el Blue Gene/L del T. J. Watson Research Center de IBM para ejecutar el programa C2, consiguiendo la simulación casi en tiempo real de la corteza cerebral de un ratón (Ananthanarayanan, Esser, Simon & Modha, 2009, p. 1). Más recientemente, en 2009, ese mismo equipo ejecutó una versión más avanzada de C2 en el Blue Gene/P del LLNL para realizar dos simulaciones. En la primera de ellas se replicó con una

resolución temporal de 1 milisegundo el funcionamiento de las 10^9 neuronas y 10^{13} sinapsis que se estima que hay en la corteza cerebral de un gato, aunque a una velocidad 643 menor que la real. La hazaña fue posible explotando al máximo las 147.456 CPUs (*central processing units*, unidades centrales de procesamiento) y los 144 TB ($144x8x10^{12}$ bits, símbolos binarios de 1 ó 0) de memoria principal de la máquina. Tan importante como el número de neuronas simuladas fue su nivel de detalle. Mientras que el perceptrón de Rosenblatt sólo replicaba la diferencia de fuerza de las conexiones sinápticas, la primera de las dos simulaciones de C2 incluía potenciales de acción, demora axonal, canales sinápticos dinámicos, conectividad tálamo-cortical y, por supuesto, aprendizaje hebbiano. Los autores del experimento compararon sus cifras con las de SyNAPSE.

SyNAPSE, siglas de *Neuromorphic Adaptive Plastic Scalable Electronics*, es un programa de DARPA que tiene por finalidad la construcción de dispositivos compactos capaces de simular redes de 10^8 neuronas y 10^{12} sinapsis. Dado que un gato tiene 7,63 veces más neuronas y 6,10 veces más sinapsis, el equipo de Ananthanarayanan superó con creces las cifras de SyNAPSE, pero con la diferencia, como hemos mencionado, de que su simulación era 643 veces más lenta que la realidad. En comparación con SyNAPSE, un cerebro humano tiene 200 veces más neuronas y 200 veces más sinapsis, por lo que el objetivo último de replicar nuestro encéfalo todavía queda muy lejos. Según Henry Markram, haría falta una máquina «20.000 veces más potente que los actuales superordenadores, con una memoria capaz de almacenar 500 veces todo el contenido del sistema actual de Internet» (Kaku, 2011, p. 138).

Markram es el director de *Blue Brain*, un proyecto que pretende replicar el cerebro humano a escala molecular aplicando técnicas de ingeniería inversa. Blue Brain consiste en un software llamado NEURON que, de momento, se ejecuta en un Blue Gene, pero en realidad el superordenador que Blue Brain necesitaría es, como dice Markram, mucho más potente. Mientras que la Blue Gene/P utilizada por el equipo de Ananthanarayanan funciona con una instalación de aire acondicionado de 6.675 toneladas que bombea 76.500 metros cúbicos de aire helado por minuto, para

refrigerar el superordenador que Blue Brain requiere habría que multiplicar esas cifras por 1.000 (Ibíd., p. 139). Tan gigantesca máquina consumiría en total 1.000 millones de vatios, lo que equivale a la producción de toda una central nuclear, una cantidad de energía monstruosa por sí sola, y más todavía si la comparamos con los insignificantes 20 vatios que gasta el cerebro humano. Podría pensarse erróneamente que Blue Brain es sólo cuestión de dinero: se invierte en multiplicar el tamaño de Blue Gene y en construirle su propia central nuclear, y ya tendríamos una verdadera IA subsimbólica de tipo humano. Pero eso sólo sería parte de la solución, porque aún faltaría descifrar el *conectoma*, es decir, el cableado de nuestro cerebro. Juntar un montón de neuronas conectadas entre sí al azar y pretender que surja una inteligencia es como lanzar cubos de pintura sobre un lienzo y esperar que aparezca un cuadro de Velázquez. Un Jackson Pollock, quizás, pero un Velázquez, no.

En cuanto a la IA subsimbólica de tipo ajeno, es la más exitosa y extendida de las dos (Santos & Duro, 2005, p. 93). Mientras que la de tipo humano no ha superado todavía el enfoque de la IA débil, es decir, que sólo sirve para progresar en las investigaciones de neurociencia, la de tipo ajeno pertenece al enfoque de la IA fuerte, pues tiene variadas y muy valiosas aplicaciones prácticas. David Rumelhart señala que la IA subsimbólica ajena presenta dos grandes diferencias respecto de la IA simbólica (Rumelhart, 1989, p. 207). Primero, que su estrategia básica consiste en utilizar una especie de neurona abstracta como unidad fundamental de procesamiento. Así, a diferencia de los programas de IA simbólica, que funcionan mediante ejecución serial, una tras otra, de miles de instrucciones por segundo, los de neuronas artificiales deben su potencia a la ejecución paralela de disparos de unas neuronas sobre otras, sin necesidad de que el número de disparos por segundo sea muy elevado. Para ilustrar la sustancial diferencia entre el procesamiento serial y el procesamiento paralelo, Rumelhart hace referencia a la *regla de los 100 pasos* (*hundred-step rule*) de Jerome Feldman. Feldman señala que el tiempo de reacción humano es de unos 500 milisegundos (Franklin, 1995, p. 144). Ése es el lapso que, según él, lleva categorizar una percepción, recuperar una memoria, desambiguar una palabra y, en general,

realizar cualquier operación cognitiva sencilla. Por otra parte, el tiempo que tarda el disparo de una neurona en producir el disparo de la siguiente es de unos 5 milisegundos, de lo cual resulta que una operación cognitiva sencilla conlleva no más de 100 activaciones seriales de neuronas. Ninguna IA simbólica es capaz de realizar nada interesante en tan sólo 100 pasos, de lo que se concluye que los 100 pasos que le toma al cerebro hacerlo deben su potencia a la distribución en paralelo de una gran cantidad de componentes.

La segunda diferencia apuntada por Rumelhart entre la IA simbólica y la IA subsimbólica es arquitectónica. Ya nos referimos a ella en el capítulo quinto a propósito de la teoría de la inteligencia de Jeff Hawkins, y es que mientras que en la IA simbólica el almacén de información y el centro de procesamiento son estructuras separadas tal y como establece la arquitectura von Neumann, en la IA subsimbólica no existe semejante distinción. El procesamiento y el almacenamiento es realizado por la misma unidad funcional: la neurona, denominada habitualmente como "unidad". El conocimiento es almacenado en las fuerzas o pesos de las conexiones interneuronales, un concepto que ya explicamos al hablar del perceptrón de Rosenblatt y que refleja la conductividad entre dos neuronas, y el procesamiento se efectúa mediante el tránsito de impulsos eléctricos a través de dichas conexiones. La información almacenada, por tanto, no está contenida de forma *explícita* en datos que pudieran ser interpretados por otro dispositivo o por un operario humano como sucede en la IA simbólica, sino que está contenida de forma *implícita*, indisociable de la estructura, determinando el modo en que ésta procesa la información entrante.

Smolensky profundiza en esta diferencia señalando que en la IA simbólica las *entidades manipuladas* por las reglas que gobiernan el sistema son también las *entidades semánticamente interpretables* (Smolensky, 1989, p. 239). Por ejemplo, símbolos como "mesa" o "silla". En cambio, en la IA subsimbólica las entidades manipuladas y las entidades semánticamente interpretables no son las mismas. Las primeras son las unidades o neuronas, mientras que las segundas son los patrones de activación de varias de esas unidades. Así, las unidades significativas, que son los

patrones, se descomponen en entidades manipulables y modificables. La modificación acontece como resultado de la experiencia mediante la variación de la fuerza de las conexiones según el principio de Hebb, al cual nos hemos referido varias veces en capítulos anteriores, y que establece, básicamente, que una conexión aumenta su fuerza cuando las neuronas unidas por ella están activas al mismo tiempo, y por el contrario puede debilitarse cuando eso no sucede. Cuando una conexión reduce su fuerza a cero, es como si físicamente desapareciese. Otros dos conceptos citados por Rumelhart y que merece la pena mencionar son los de *fan-in* y *fan-out*. El fan-in de una unidad es el número de unidades que proyectan sobre ella, ya sea excitándola o inhibiéndola, y el fan-out de una unidad es, a la inversa, el número de unidades sobre las que ella proyecta. En el sistema nervioso de los seres humanos hay neuronas con un fan-in del orden de 150.000 (Kandel, Schwartz & Jessell, 1995, p. 27), mientras que en los sistemas conexionistas la cifra es, por lo general, mucho menor.

Siendo éstas las características principales de las redes de neuronas, Rumelhart enumera una serie de problemas habituales en los modelos de la cognición para los que las redes de neuronas ofrecen soluciones eficaces (Rumelhart, 1989, p. 216). Éstos son: problemas de satisfacción de restricciones (*constraint-satisfaction problems*); búsqueda de la mejor coincidencia (*best-match search*), reconocimiento de patrones y memoria direccionable por el contenido; implementación automática de generalizaciones basadas en la similaridad; mecanismos simples y generales de aprendizaje adaptativo; y degradación elegante (*graceful degradation*) ante la sobrecarga o el daño de la información.

Los *problemas de satisfacción de restricciones* son aquellos cuya solución es dada por la satisfacción de un gran número de restricciones interrelacionadas. Para este tipo de problemas, las redes de neuronas se conceptualizan como *redes de restricciones* (*constraint networks*) en las que cada unidad o neurona representa una hipótesis, y cada conexión representa una restricción entre las hipótesis. Por ejemplo, supongamos un cubo de Necker, que es ese cubo en perspectiva isométrica que no se sabe si se está viendo desde arriba o desde abajo porque todas sus aristas son

igualmente visibles. La solución para dirimir desde dónde se está viendo el cubo pasaría por definir dos redes de neuronas con los vértices del cubo como unidades. En la red A las unidades representan hipótesis que en su conjunto describen el cubo como visto desde arriba, mientras que en la red B las unidades representan hipótesis que en su conjunto describen el cubo como visto desde abajo. Cada unidad de cada red, o lo que es lo mismo, cada vértice, recibe conexiones excitatorias de fuerza 2 de cada uno de sus tres vértices adyacentes, pues todo vértice de un cubo tiene tres vértices adyacentes, mientras que recibe conexiones inhibitorias de fuerza 3 de dos unidades de la otra red que representan hipótesis incompatibles. Por ejemplo, en una red un vértice puede ser interpretado por la hipótesis de que pertenece a la cara más cercana al espectador, mientras que en la otra red el mismo vértice puede ser interpretado como perteneciente a la cara más lejana al espectador.

Si una red de este tipo, dice Rumelhart, se pone en marcha, alcanzará un estado óptimo de equilibrio en el que se satisface el mayor número posible de constricciones, es decir, de relaciones de hipótesis, dando prioridad a las constricciones más fuertes. El procedimiento a través del cual el sistema se *establece* (*settles*) en dicho estado se denomina *relajación* (*relaxation*) (Ibíd., p. 216). En el caso particular del cubo de Necker la conclusión es que no se puede decidir desde dónde está siendo observado, pues tan fuertes son en su conjunto las hipótesis de que está siendo visto desde arriba como las de que está siendo visto desde abajo. Serían dos estados o interpretaciones igualmente fuertes, pues lo que cada estado representa es una interpretación global surgida de las hipótesis.

Los problemas de *búsqueda de la mejor coincidencia*, *reconocimiento de patrones* y *memoria direccionable por el contenido* son todos ellos, dice Rumelhart, variantes del problema general de la mejor coincidencia, el cual consiste en encontrar la memoria almacenada que más se parezca a la entrada de información. Para este propósito es útil distinguir las unidades de la red en dos tipos: *visibles*, que corresponden a los contenidos almacenados en la red en el sentido de que cada patrón almacenado es un posible patrón de activación de esas unidades; y *ocultas* (*hidden*),

que corresponden a las propiedades estructurales compartidas de los patrones almacenados y que participan en los procesos de almacenamiento y recuperación. La recuperación se efectúa mediante la asignación de los valores de un patrón a *algunas* de las unidades visibles, y dejar que la red actúe hasta que complete los valores del resto de las unidades visibles.

Empleando diversas reglas de aprendizaje se consiguen redes con las siguientes cinco propiedades, dice Rumelhart (Ibíd., p. 222): respuesta de reconocimiento ante patrones familiares, respuesta de extrañeza ante patrones no familiares, memoria direccionable por el contenido, respuesta de asimilación y respuesta a prototipos. Las dos primeras no necesitan explicación: la red responde a patrones entrantes familiares y no familiares de distinta manera. La *memoria direccionable por el contenido* es muy útil, ya que basta con disponer de un fragmento de un patrón para recuperarlo en su totalidad. Gracias a esta cualidad, los sistemas conexionistas solucionan de manera automática el *problema de la selección del marco*, al que nos referimos en el capítulo anterior, y que es irresoluble por la IA simbólica como veremos en el capítulo siguiente. Al procesar la información entrante a través de las conexiones establecidas como resultado de las experiencias anteriores, las redes de neuronas están aplicando de ese modo un marco o esquema organizador.

Smolensky utiliza el mismo ejemplo del reconocimiento de los objetos de una habitación utilizado por Minsky para dar cuenta de la solución del problema de la selección del marco desde una perspectiva simbólica en su artículo antes citado *A framework for representing knowledge*. El proceso de codificación subsimbólica del marco de una habitación comienza por describir unas cuantas habitaciones imaginarias utilizando 40 características, tales como "tiene techo", "tiene ventana", "tiene un lavabo" (Smolensky, 1989, p. 248). A continuación se construye una red con un nodo para cada característica, y se entrena para que sus nodos reflejen la frecuencia estadística con la que estas características suelen presentarse simultáneamente. Así, la identificación de unas cuantas partes, como que hay una cafetera, ayuda a identificar la totalidad, pues lo habitual es que una cafetera esté en la cocina o en el salón pero

no en un cuarto de baño, y recíprocamente la totalidad ayuda a identificar otras partes, pues sobre la conjetura de que estamos en una cocina se anticipa que probablemente habrá también una nevera, una lavadora y un horno. En este sistema, a diferencia del propuesto por Minsky, no hay unidades que representen los marcos o totalidades, en este caso tipos de habitaciones, sino que una cocina es un patrón de activación, una información distribuida que, aunque no está realmente ahí al nivel más fundamental, sí está presente en forma de propiedad de una descripción de alto nivel que surge a partir de inferencias estadísticas de sus partes. Y, a la inversa, profundizando hacia niveles más bajos, las partes de los que se componen los marcos, como la nevera, la lavadora y el horno, son también descomponibles (Rumelhart, McClelland & The PDP Research Group, 1986, p. 255).

En cuanto a la *respuesta de asimilación* se refiere a que cuando la red recibe un patrón entrante ligeramente distinto al almacenado, el sistema lo rectifica para que se parezca a este último. Y, finalmente, la *respuesta a prototipos* quiere decir que la red es capaz de responder a patrones que nunca hayan sido experimentados, siempre y cuando guarden un parecido con el prototipo que la memoria alberga como resultado de la exposición a varios patrones similares. Un modelo de la inteligencia con estas cinco propiedades no tendría problemas para elicitar memorias pertinentes en tanto que similares. Así, no haría falta un sistema de etiquetado e indización de memorias tan complejo y, por otra parte, defectuoso, como el que vimos en el capítulo anterior que Schank se veía obligado a proponer por su planteamiento simbólico.

No obstante, a pesar de sus ventajas, el reconocimiento de patrones mediante redes de neuronas no se halla exento de dificultades. Hubert Dreyfus comenta el caso de una red utilizada por los militares para identificar tanques ocultos en un bosque (Dreyfus, 1992, p. xxxvi). Suministraron al sistema una serie de fotos A de tanques en un bosque, y después otra serie de fotos B tomadas *otro día* del mismo bosque sin tanques. Mediante la respuesta a la extrañeza la máquina demostró ser capaz de discriminar ambos conjuntos de fotografías. Sin embargo, para asegurarse de que todo había salido bien, los investigadores tomaron más fotografías en una tanda C, y para su

decepción resultó que la máquina era incapaz de discriminarlas. Tras muchas hipótesis, alguien se percató de que las primeras fotografías A con tanques habían sido tomadas en un día soleado, mientras que las de sin tanques B correspondían a un día con nubes. En resumen, lo que la red había aprendido a diferenciar era fotografías con sombras (día soleado) y sin sombras (día nublado). Ciertamente, este caso, como señala Dreyfus, corresponde a los inicios de las redes de neuronas, pero sirve para ilustrar algunas de las dificultades que entraña esta tecnología.

La tercera gran virtud de los sistemas de IA subsimbólica es la *generalización automática basada en la similaridad*. Los programas de IA simbólica son muy eficaces en la resolución de aquellos problemas para los que han sido diseñados, pero pueden llegar a producir resultados desastrosos ante situaciones nuevas. Las situaciones nuevas son justamente un entorno en el que las redes de neuronas responden bien. Dado que las similaridades entre patrones están representadas junto con los patrones en los pesos de las conexiones, la generalización basada en la similaridad es una propiedad automática de los modelos conexionistas (Rumelhart, 1989, p. 223). La clave para conseguir buenas generalizaciones es un proceso de aprendizaje adecuado.

Respecto del *aprendizaje* de las redes de neuronas, ya señalamos antes que Rosenblatt descubrió una regla de aprendizaje para las redes de dos capas, como su perceptrón, pero no para las multicapa, que son las que poseen además de las capas de entrada y de salida una tercera capa de unidades ocultas. Las redes de dos capas, apunta Rumelhart, son eficaces para algunos tipos de procesos, mientras que para otros no lo son, debido a que, al no tener unidades ocultas, carecen de representaciones internas. «Cuando la representación proporcionada por el mundo exterior es tal que la estructura de similaridad entre los patrones de entrada y salida es muy diferente, una red sin representaciones internas (esto es, una red sin unidades ocultas) será incapaz de trazar las correspondencias (*mappings*) necesarias» (Ibíd., p. 223). Un caso de operación no realizable por las redes de dos capas pero sí por las multicapa es la disyunción exclusiva, representada por el operador lógico XOR. El problema de las redes multicapa en los años 60 era que, como observaron Minsky y

Papert en su libro contra el perceptrón de Rosenblatt, no había una regla de aprendizaje para ellas que fuese tan poderosa como la *regla delta* descubierta por Bernard Widrow y Ted Hoff (Ibíd., p. 214). La regla delta dice: $g(a_i,\tau_i)=\varepsilon\cdot(\tau_i-a_i)$. El primer miembro de la ecuación es una función g que relaciona la activación de una unidad *i* (a_i) con una variable de aprendizaje τ. El segundo miembro está formado por la constante de proporcionalidad ε que representa la tasa de aprendizaje y una sustracción de los dos elementos anteriores.

La principal contribución de Rumelhart, McClelland y el grupo PDP a la IA subsimbólica consistió en el descubrimiento a mediados de los 80 de una regla de aprendizaje para las redes multicapa: la *regla delta generalizada*. Ésta consta de dos fases. En la primera se le aplica una entrada a la red, y en la segunda se compara la salida producida con el resultado deseado. La diferencia es denominada *error*. Si el error es nulo, la red está bien ajustada. De lo contrario, el error es enviado hacia atrás en la jerarquía, en dirección a la capa de entrada, para modificar los pesos de las conexiones de entrada de las unidades de la segunda capa, y así sucesivamente hacia capas anteriores si las hubiera. Este proceso se denomina *retropropagación* o *propagación hacia atrás del error* (*backpropagation of error*). Una vez se han hecho los ajustes pertinentes en base al error, se introduce una nueva entrada en la red y se repite el ciclo hasta que el error sea nulo o aceptablemente bajo. Este sistema, dice Rumelhart, funciona para cualquier tipo de red (Ibíd., p. 226).

La quinta y última cualidad distintiva de las redes de neuronas enumerada por Rumelhart es la *degradación elegante*. Debido a que en una red todas las unidades participan en el almacenamiento de muchos patrones y que cada patrón implica a varias unidades diferentes, la pérdida de algunos componentes degrada la información almacenada, pero no la destruye. Esto contrasta radicalmente con las memorias convencionales de los ordenadores. Éstas consisten en secuencias binarias, siendo cada dígito almacenado en un solo lugar, de manera que si ese lugar resulta dañado, es imposible recuperar el dígito, lo cual puede tener el efecto devastador de que todo un programa enorme deje de funcionar a causa de tan ínfima pérdida.

Como vemos, la IA subsimbólica en la actualidad es un programa de investigación muy prometedor en el que hay puestos muchos esfuerzos y financiación, tanto en la vertiente de la IA humana en la línea de Blue Brain, como en el de la IA ajena, a la que pertenecen SyNAPSE y las redes estudiadas por Rumelhart y compañía. No obstante, aun suponiendo que la IA subsimbólica superase las dificultades técnicas de las que nos ocuparemos en el próximo capítulo, sus posibilidades de divulgación comercial seguirían siendo escasas por la misma razón que descubrió Marvin Minsky cuando era joven: porque para crear verdaderas inteligencias artificiales con redes de neuronas artificiales hacen falta máquinas que, por su coste y su tamaño, sólo están al alcance de unos pocos, como las agencias militares.

Evolución artificial

José Santos y Richard Duro, expertos en robótica, definen la evolución artificial como la «simulación en un ordenador del mismo procedimiento que ha tenido lugar a lo largo de millones de años en el mundo natural» (Santos & Duro, 2005, p. 37). Desde el punto de vista histórico, se trata de una técnica surgida a finales de los años 80 en diversas universidades de Inglaterra y Estados Unidos con el objetivo de automatizar el diseño de sistemas autónomos, capaces de sobrevivir con la menor intervención humana posible. La evolución artificial es aplicable a las dos grandes aproximaciones a la arquitectura cognitiva que hay en robótica: la basada en conocimiento y la basada en comportamiento (Brooks, 1991, p. 402). La *arquitectura basada en conocimiento* es la tradicional, heredada de la IA simbólica, que consiste en descomponer la cognición en funciones sucesivas que procesan la información entrante hasta producir la salida. Por ejemplo, las entradas de los sensores proyectarían la información a un módulo de interpretación de datos sensoriales, que posteriormente pasarían sus resultados a un modelo del entorno, que pasaría sus resultados a un módulo de planificación, que finalmente pasaría sus resultados a un módulo de ejecución que realizaría una conducta de salida en respuesta a la entrada original. De esta forma, con que fallase

uno solo de los módulos, se interrumpiría el flujo de información, y fallaría el sistema entero. En cambio, en la *arquitectura basada en comportamiento* la cognición se divide en módulos paralelos que son relativamente independientes entre sí, en tanto que todos tienen acceso a las entradas de los sensores y a las salidas de los actuadores, por lo que si alguno de ellos falla, lo habitual es que el resto pueda seguir operando. Ejemplos de esos módulos serían los de evitar obstáculos, vagar, explorar, construir mapas, chequear cambios o identificar objetos. Como decimos, ambas arquitecturas, la basada en comportamiento y la basada en conocimiento, pueden ser desarrolladas por evolución artificial. Sin embargo, en la actualidad la más utilizada es la basada en comportamiento, o bien un enfoque híbrido, mientras que la basada sólo en conocimiento está en declive desde los años 90, dicen Santos y Duro, porque implica problemas tan graves como el del marco (Santos & Duro, 2005, p. 14).

Volviendo a la definición de la evolución artificial, la primera diferencia que encontramos entre la evolución natural y la artificial es que esta última procede *simulando* de manera virtual a los individuos y a sus entornos. Lo ideal, señalan Santos y Duro, sería evolucionar individuos reales en forma de robots que interactuasen con entornos reales, pero esta estrategia es impracticable debido a que recrear la evolución de poblaciones de cientos de individuos a lo largo de miles de generaciones llevaría años de trabajo si se hiciese con robots reales. Por tanto, la simulación acelerada por computadora se impone como la única estrategia viable, lo cual implica una serie de problemas que más adelante comentaremos.

La segunda diferencia más ostensible entre la evolución natural y la artificial reside en que la natural procede de manera ciega, sin un objetivo. Santos y Duro citan al premio Nobel de medicina François Jacob, quien decía: «La selección natural no trabaja como un ingeniero, sino como un chapucero, un chapucero que todavía no sabe qué va a producir [...] Un chapucero que aprovecha todo lo que encuentra para obtener algún objeto útil» (Ibíd., p. XII). En cambio, la evolución artificial sí procede hacia una *dirección*, concretamente la determinada por los investigadores. La dirección se determina mediante la definición de la calidad de los individuos. La *calidad* consiste

en una cantidad numérica que será mayor cuanto mejor realice un individuo durante su tiempo de vida las tareas deseadas por los investigadores. Por ejemplo, si lo que se desea es fabricar aspiradoras inteligentes, de esas que recorren las casas limpiando sin chocar con las paredes, por cada choque que cometa un individuo en el simulador se le restarán puntos de calidad, y por la velocidad con la que realice su cometido se le sumarán. Si por azar en el proceso evolutivo surgiera un individuo que no realizase bien las tareas propias de una aspiradora pero, en cambio, por tuviese la asombrosa habilidad de dibujar paisajes con el rastro dejado por sus escobillas, recibiría sin embargo una mala nota de calidad, porque la evolución artificial está, como decimos, dirigida hacia la obtención de un cierto tipo de individuos.

La asignación de calidad puede realizarse desde dos perspectivas: local y global (Ibíd., p. 74). La *local* otorga puntos de calidad a cada acción particular, mientras que la *global* puntúa al individuo al final de su vida en función de lo bueno que fue, en términos generales, en la realización de su objetivo. Ambas perspectivas presentan problemas. Por un lado, en la local es muy difícil para el diseñador determinar el valor de cada acción particular para la consecución del objetivo final, pues la misma acción puede ser buena o mala en función de las circunstancias. Por el otro, en la global lo difícil es repartir el crédito entre las diversas acciones para determinar cuánto contribuyó cada una. Dicho de otro modo, el principal problema de la perspectiva global es que podría ocurrir que un individuo fuese el mejor de su generación, pero sin embargo se comportase como la paloma de Skinner, que realizaba un montón de conductas superfluas porque no sabía discriminar cuáles eran necesarias y cuáles no para conseguir el alimento. Dentro de las asignaciones globales cabe distinguir entre las externas y las internas. En las *externas* es un evaluador ajeno al individuo el que, desde una posición mejor informada, asigna la calidad. En cambio, en las *internas* el propio individuo sabe lo bien o mal que ha realizado la tarea gracias a un indicador, frecuentemente denominado "energía interna". El nivel de energía sube o baja en función de parámetros como la cantidad de veces que se haya realizado la tarea, lo que se haya tardado o el tiempo transcurrido desde la última realización.

Una vez asignada la calidad, ésta sirve para determinar qué individuos se reproducirán. A mayor calidad, más probabilidades tiene un individuo de pasar su información genética a la siguiente generación. La decisión de qué sujetos tienen descendencia se realiza mediante un operador de *selección*, el cual, sin entrar en detalles, puede consistir en una combinación de distintos métodos, como el de la ruleta, la selección por torneo o el elitismo (Ibíd., p. 49). Junto con la selección, los otros dos operadores básicos presentes en los distintos métodos de evolución artificial son el cruce y la mutación. El operador de *cruce* combina el material genético de dos o más progenitores para producir un descendiente, mientras que el operador de *mutación* cambia el contenido del material genético de un determinado cromosoma. Utilizando una terminología análoga a la aplicable a los seres vivos, se dice que la información genética de los seres sometidos a un proceso de evolución artificial se codifica en genes, los cuales se organizan en cromosomas, que en su conjunto componen el genotipo del individuo.

De entre los diversos algoritmos utilizados para recrear el proceso de evolución artificial, Santos y Duro señalan los siguientes: algoritmos genéticos, estrategias evolutivas, programación genética, programación evolutiva y coevolución (Ibíd., p. 53). Los *algoritmos genéticos* (AGs) se caracterizan porque el factor dominante para el propósito de la evolución es el operador de cruce. Un problema muy importante de los AGs es el de la *engañosidad* (*deceptivity*), que se produce cuando dos progenitores con genes de alta calidad dan lugar a un descendiente de baja calidad. Una de las posibles explicaciones de este fenómeno es la *epístasis*, concepto similar al de totalidad (*wholeness*) que vimos en el capítulo anterior al comentar la teoría de la inteligencia de Douglas Detterman, y que refiere a la interdependencia de la calidad de varios genes, de tal manera que la variación de uno solo de ellos puede dar lugar a que el resto no produzca buenos resultados. Es el caso, por ejemplo, de las redes de neuronas artificiales, en las que la alteración del peso de una sola unidad puede producir el mal funcionamiento de toda la red. Así, los progenitores pueden tener en sus genes codificados sendos juegos de pesos iniciales de sus redes de neuronas que sean

exitosos, pero al combinarlos el juego resultante de pesos de su descendiente puede ser pésimo, e incluso inoperante. Cuanto mayor es el nivel de epístasis de un genotipo, de mayor tamaño serán los *bloques* o *esquemas* de genes interrelacionados a manipular, y por tanto menos posibilidades combinatorias habrá.

En cuanto a las *estrategias evolutivas* (EE), su principal diferencia respecto de los AGs, dicen Santos y Duro, es que el operador que guía el proceso evolutivo no es el cruce, sino la mutación. En la práctica, sin embargo, no suelen emplearse AGs o EEs puros, sino una combinación. Por su parte, la programación genética (PG) y la programación evolutiva (PE) son las respectivas contrapartidas en lenguajes de alto nivel de los AGs y las EEs. Es decir, que en la *programación genética* el objeto de evolución es un programa definido en un lenguaje de alto nivel al que se le aplican operadores de cruce, mientras que en la *programación evolutiva* se le aplican operadores de mutación a un programa también definido en un lenguaje de alto nivel.

Por último, la *coevolución* consiste en la evolución simultánea de dos o más genotipos que corresponden a problemas diferentes. La coevolución puede ser *cooperativa*, cuando las subpoblaciones de individuos cooperan, y *competitiva*, cuando compiten. En el caso de la coevolución competitiva acontece un fenómeno muy interesante denominado *efecto de la Reina Roja*. Su nombre es un homenaje al fragmento de la novela de Lewis Carroll *A través del espejo* en el que Alicia y la Reina Roja corren, pero por más que corren no consiguen avanzar porque todo lo que las rodea se mueve en la misma dirección que ellas. Análogamente, en la coevolución competitiva, si suponemos un grupo de cazadores y otro de presas, no existe una medida de calidad absoluta, pues la calidad de las presas es relativa a la calidad de los cazadores, así como el avance de Alicia y la Reina es relativo al avance del entorno. De esta forma, un depredador muy exitoso en una generación puede ser de los peores si se lo inserta en otra generación en la que las presas sean más escurridizas. En el ámbito de la inteligencia sucede lo mismo, tal y como vimos en el capítulo quinto a propósito del efecto Flynn: no se puede comparar los resultados de los tests de inteligencia de sujetos de diferentes épocas, porque el contexto cambia.

Hasta aquí hemos visto cómo procede la evolución artificial para automatizar el diseño de los individuos en busca de los mejores genotipos para la realización de una tarea. Lo que se puede evolucionar, y por tanto codificar en dichos genotipos, son: los controladores de comportamiento, el hardware o morfología del individuo y el modelo del entorno (Ibíd., p. 81). Empezando por los *controladores de comportamiento*, son las estructuras que soportan los mecanismos cognitivos del sujeto, y pueden ser de tres tipos: programas de control explícito, expresiones matemáticas y la definición de una estructura de procesamiento. Santos y Duro se centran en las estructuras de procesamiento debido a que son las más utilizadas en robótica de comportamiento, de la que ya hemos dicho que es más exitosa que la basada sólo en conocimiento. Las principales estructuras de procesamiento son los sistemas de clasificación, los sistemas de lógica borrosa y las redes neuronales artificiales. Los *sistemas de clasificación* son una especie de sistemas de producción. La *lógica borrosa* (*fuzzy logic*), a grandes rasgos, se diferencia de la lógica convencional en que la pertenencia de un elemento a una determinada clase no es booleana, de todo o nada, sino gradual. Respecto a las redes de neuronas artificiales (RNA), son los controladores de comportamientos más utilizados, razón por la cual hemos dispuesto este tema de la evolución artificial justo a continuación del dedicado al estado actual de la IA subsimbólica.

Los sistemas conexionistas, dicen Santos y Duro, son preferidos por su capacidad de aprendizaje, el procesamiento en paralelo y la alta tolerancia a fallos. El *aprendizaje* puede realizarse de manera supervisada o no supervisada. Un método *supervisado* sería la antes mencionada *regla delta generalizada*, mientras que un ejemplo de *no supervisado* sería el *aprendizaje Q*, basado en el condicionamiento operante, por refuerzo y castigo (Ibíd., p. 129). Obviamente, los métodos no supervisados son los más usados, dado que no requieren de la intervención de un operario humano. Los genes que codifican una red de neuronas artificial suelen contener parámetros tales como los pesos, el número de capas, unidades por capa y el tipo de unidades, e incluso otras características más refinadas, como el efecto de los neuromoduladores del tipo del óxido nítrico (Ibíd., p. 186).

Respecto a la evolución del *hardware o morfología* del individuo, los genes permiten codificarlo todo, desde la circuitería sobre la que se ejecutan las estructuras de procesamiento hasta la forma del cuerpo. Centrándonos en la morfología del cuerpo, el gran problema de la evolución del hardware es que, al realizarse de manera simulada, la posterior transferencia a sistemas reales se vuelve compleja o difícil. En este campo el investigador Karl Sims ha obtenido mediante evolución artificial una impresionante colección de criaturas virtuales que, a través de sucesivas generaciones, han terminado por desarrollar morfologías adecuadas para realizar ciertas tareas. Por ejemplo, en la lucha competitiva por atrapar objetos, algunas han desarrollado una especie de pinzas, y para nadar, otras han desarrollado aletas y cuerpos flexibles como los de los peces. El traslado exitoso de estos diseños virtuales a robots reales depende en buena medida del realismo con el que se hayan recreado en el mundo virtual de la computadora las características del mundo físico, tales como las fuerzas inerciales o el rozamiento, pues si el mundo virtual difiere demasiado del real, las morfologías obtenidas evolutivamente resultarán poco o nada eficaces.

Por último, los genes también sirven para codificar, y por tanto para evolucionar, un *modelo del entorno*. En función de la riqueza del modelo del entorno que tenga un individuo, su comportamiento se dice que es más reactivo o más deliberativo (Ibíd., p. 24). Si posee un modelo pobre o directamente ningún modelo, su comportamiento será más *reactivo*. Por el contrario, cuanto más detallado sea el modelo, más *deliberativa* podrá ser su conducta, en tanto que dispondrá de una mayor capacidad para deliberar, es decir, planificar a través del tiempo y del espacio. Sobre este tema, Santos y Duro lamentan la escasez de propuestas alternativas a la suya propia. El modelo del entorno, dicen, se divide en un *modelo del mundo* y un *modelo interno*: «El primero es una representación que define cómo es la percepción del mundo por parte del robot, al predecir cómo va a ser la percepción siguiente en función de la percepción externa anterior y tras realizar una determinada acción. Por su parte, el estado interno define cómo se van a satisfacer las motivaciones en función de la acción realizada y de la percepción de satisfacción de motivaciones interna del

robot» (Ibíd., p. 105). Ambos modelos, del mundo y del interior, evolucionan a través un mecanismo cognitivo darwinista que Santos y Duro han elaborado inspirándose en la *teoría de la selección de grupos de neuronas* (*theory of neural group selection*, *TNGS*) del premio Nobel de química Gerald Edelman. Santos y Duro han bautizado a su mecanismo como *cerebro darwinista multinivel* (*multilevel darwinist brain*, MDB).

En el MDB existen dos niveles: inconsciente y consciente. En el inconsciente el robot evoluciona poblaciones de modelos del mundo y de modelos internos. La función de calidad que determina cuáles son los modelos que pasarán sus características a la siguiente generación consiste en una evaluación del éxito de sus predicciones, un criterio en la línea de la metáfora de Jeff Hawkins de la inteligencia como máquina del tiempo que vimos en el capítulo anterior. Así, los modelos que generan predicciones exitosas evolucionan, mientras que los que generan predicciones fallidas desaparecen para dejar sitio a cruces y mutaciones de los anteriores. Cada vez que existe la necesidad de realizar una acción o determinar una estrategia, dicen Santos y Duro, el robot pasa al nivel consciente, en el que se utiliza el mejor modelo del mundo y el mejor modelo interior disponible en ese momento para evolucionar sobre ellos «posibles estrategias hasta que obtenga una estrategia que maximiza la consecución de sus motivaciones» (Ibíd., p. 107). Por tanto, la evolución del modelo del entorno consiste en una evolución dentro de la evolución, es decir: una evolución a escala ontogenética, o individual, de un sujeto que participa en la evolución a escala filogenética, o de la especie. Una vez que un individuo llega al final de su vida y ha sido seleccionado para reproducirse, caben dos posibilidades respecto del modelo del entorno exitoso que ha desarrollado: pasarlo o no pasarlo a su descendencia.

Si los investigadores deciden pasarlo, entonces estarían adoptando un enfoque *lamarquista* de la evolución (Ibíd., p. 142). El biólogo francés Jean Baptiste de Monet, más conocido por su título nobiliario de Chevalier de Lamarck, propuso a finales del siglo XVIII dos leyes de la evolución. La primera es la *ley del uso y el desuso*, según la cual las partes del cuerpo que se usan repetidamente, se desarrollan, y por el contrario, las que no se usan, se atrofian y finalmente desaparecen. La segunda es la

ley de la herencia, que postula que los animales transmiten a sus descendientes los caracteres que han adquirido a lo largo de su vida. Los biólogos posteriores a Lamarck demostraron que esta segunda ley es falsa, a pesar de lo cual en la evolución artificial sí puede hacerse verdadera, pues no hay nada que impida a los investigadores programar a sus criaturas para que transmitan hereditariamente lo que han aprendido. En el caso de las redes de neuronas, por ejemplo, en vez de introducir en el proceso de cruce y mutación el conjunto de valores de los pesos de las neuronas con el que inició su vida, introduciría el conjunto final de valores de los pesos modelados por la experiencia. Adoptar este enfoque lamarquista tiene ventajas e inconvenientes. La ventaja más evidente es que la nueva generación empezará su vida disponiendo de un modelo del entorno derivado del modelo de progenitores exitosos, por lo que sus primeras acciones, salvando el problema de la engañosidad, serán mucho más diestras que las de sus predecesores. Sin embargo, el lamarquismo presenta un gran inconveniente, y es que reduce rápidamente la *diversidad genética* de la población.

Para entender la importancia de la diversidad genética debemos entender primero lo que es una hipersuperficie de calidad. Imaginemos una especie animal con un genotipo consistente en un único cromosoma de tan sólo 2 genes, X e Y, cada uno de los cuales pudiera tener un valor de entre 0 y 100. Dibujemos un eje cartesiano con los 100 posibles valores del gen X en la horizontal y los 100 posibles valores del gen Y en la vertical. Habría por tanto 100^2 individuos posibles, o lo que es igual, 10.000 genotipos diferentes. Cada uno de ellos daría lugar a un individuo al que se le adjudicaría, tras su tiempo de vida, un valor de calidad de entre 0 y 10, que podemos representar en el eje Z. Así, cuanto mayor es el relieve Z de un punto (X,Y), mayor es la calidad del individuo codificado por ese genotipo. Este mapa, con la calidad en una de sus dimensiones, es lo que se denomina *hipersuperficie de calidad* (*fitness landscape*) (Ibíd., p. 41). Pues bien, al principio de todo proceso evolutivo artificial hay varios individuos generados aleatoriamente, y en consecuencia dispersos por toda la hipersuperficie. El inconveniente de aplicar una estrategia lamarquista es que los mejor dotados copan la variedad genotípica con gran celeridad, restringiendo la

búsqueda de nuevos individuos a las coordenadas que están en sus inmediaciones, y por tanto descartando el resto de la hipersuperficie. De esa manera, la búsqueda evolutiva podría limitarse a un área, digamos por ejemplo (27±5,60±5), en la que el pico máximo de calidad fuera de 6. A la larga, esa estrategia lamarquista llevaría a la obtención de un genotipo de, como mucho, calidad 6, mientras que los genotipos de calidad 10, que son los que el investigador aspira a encontrar, permanecerían ocultos en alguna otra región de la hipersuperficie. Por supuesto, la hipersuperficie del ejemplo que hemos puesto es explorable en su totalidad, pues no hay problema en simular el comportamiento de 10.000 individuos para evaluar su calidad, pero en la realidad los genotipos son mucho más complejos, con miles de genes que dan lugar a posibilidades combinatorias astronómicas imposibles de computar una por una.

La otra posibilidad, la de no transferir a la descendencia la experiencia adquirida, es, al contrario que el lamarquismo, una estrategia que no disminuye la diversidad genética, y por tanto permite encontrar máximos de calidad a lo largo de toda la hipersuperficie (Ibíd., p. 146). Se parte del hecho de que un individuo aprende más rápido cuanto menor sea la diferencia entre la configuración inicial determinada por su genotipo con la que viene al mundo y la configuración final aprendida. Al estar la calidad determinada en parte por la velocidad del aprendizaje, los individuos que más rápido aprendan los mejores modelos del entorno se reproducirán más. De este modo, las sucesivas generaciones irán reduciendo progresivamente la distancia entre las configuraciones iniciales y las configuraciones finales aprendidas que representan buenos modelos del entorno, hasta llegar un momento en el cual las configuraciones iniciales, esto es, las codificadas en el genotipo, contendrán directamente buenos modelos del entorno. Y no sólo buenos modelos del entorno, sino cualquier rasgo susceptible de ser evolucionado y heredado, incluyendo las estructuras de control, el hardware y la morfología. Esta transferencia de los rasgos adquiridos por los individuos al genotipo de la especie de manera indirecta, sin herencia lamarquista, se conoce como el *efecto Baldwin*, por haber sido propuesto a finales del siglo XIX por el biólogo norteamericano James Baldwin. El efecto Baldwin explica por qué las cigüeñas, de las

que hablamos en el capítulo tercero a modo de ejemplo, han desarrollado evolutivamente la habilidad innata de construir nidos, así como también explica por qué los seres humanos tenemos una predisposición innata para el lenguaje. Los individuos que tienen un genotipo más cercano a la expresión innata de las características más ventajosas para la supervivencia son los que más rápido las adquieren, y por tanto los que más se reproducen.

Para que este repaso a grandes rasgos del estado actual de las técnicas de evolución artificial quede completo debemos dar cuenta de las diferentes *arquitecturas de interconexión* posibles. La robótica basada en comportamiento admite dos arquitecturas de interconexión, dicen Santos y Duro (Ibíd., p. 172). Por un lado, la *monolítica*, que consiste en codificar el comportamiento global en un único módulo, y por el otro, la *modular*, que descompone el comportamiento global en comportamientos más simples. Obviamente, la tendencia actual en robótica es la modular, como también lo es en psicología y neurociencia. A su vez, la arquitectura de interconexión modular puede ser jerárquica o distribuida. Las arquitecturas *jerárquicas* o *centralizadas* distinguen ciertos comportamientos de alto nivel que deciden la activación de los de bajo nivel. Por el contrario, en las *distribuidas* no existen jerarquías, sino que todos los módulos compiten a cada instante por hacerse con el control de los actuadores mediante algún tipo de arbitraje o de un sistema de inhibiciones entre ellos.

Un ejemplo de arquitectura modular jerárquica sería la *arquitectura subsumida* de Rodney Brooks (Brooks, 1991, p. 408), quien a finales de la década de los 80 fue el pionero impulsor del enfoque basado en comportamiento frente al tradicional basado en conocimiento. Las arquitecturas modulares jerarquizadas, como la de Brooks, presentan dos ventajas principales: que los comportamientos pueden ser desarrollados evolutivamente de manera independiente, y que los comportamientos de nivel inferior pueden ser reutilizados por otros nuevos de nivel superior. Sin embargo, tienen el inconveniente de que la descomposición del comportamiento más complejo en otros más simples no siempre está clara, y suele depender no del proceso

evolutivo como sería deseable, sino de decisiones de diseño tomadas por los investigadores, unas decisiones no siempre acertadas porque no siempre pueden prever lo que sucederá a lo largo de la evolución. En cuanto a las arquitecturas modulares distribuidas, su uso es poco frecuente, señalan Santos y Duro, debido a que en ellas es difícil expandir el diseño existente mediante la adición de nuevos comportamientos. La causa es similar a la que ya hemos visto que dificulta la adición de reglas a los sistemas de producción, y es que, al tratarse de sistemas horizontales en los que todas las reglas o, en este caso, comportamientos, tienen el mismo rango, el añadido de tan solo un nuevo elemento puede modificar radicalmente de manera catastrófica el funcionamiento del sistema entero. En definitiva, la opinión de Santos y Duro acerca de las arquitecturas de interconexión es que: «Se utilice una u otra aproximación consideramos que conmutar entre comportamientos no es la forma más natural de actuar. Generalmente, en los seres naturales, están presentes varios comportamientos al mismo tiempo, dependiendo de las necesidades individuales, objetivos y deseos. Los comportamientos comúnmente interactúan, modulándose entre ellos» (Santos & Duro, 2005, p. 181).

Siendo la evolución artificial tal y como la hemos descrito en líneas generales, se trata de una técnica que, aunque ha evolucionado mucho desde su nacimiento hace algo más de dos décadas, todavía está dando sus primeros pasos. Prueba de ello es la simpleza de los seres robóticos producidos por ella. Entre los más destacados podemos nombrar las antes mencionadas aspiradoras autónomas que limpian el suelo sin chocar contra las paredes, los robots de Lego jugadores de fútbol que participan en las competiciones de la Robot League, y el famoso perro robótico Aibo diseñado por Sony, el cual se vendía a un precio aproximado de 2.500 dólares hasta el año 2006, fecha en la que se canceló su comercialización. Si examinamos de cerca el ingente trabajo de ingeniería que implican estas máquinas, es de justicia reconocer que se trata de inventos de alta tecnología, obra de investigadores muy talentosos, algunos de los cuales son españoles por cierto, como José Santos y Richard Duro, de la Universidad de La Coruña. Pero si los comparamos con la inteligencia de un ser humano, cabe

preguntarse si la evolución artificial es un camino que pueda conducir en un tiempo razonable a la obtención de verdaderas inteligencias artificiales en sentido fuerte. Ésta es una cuestión que abordaremos en el próximo capítulo.

6.3. La IA simbólica en la actualidad

El sentido común (*bon sens*), decía Descartes, es la cosa mejor repartida del mundo (Descartes, *Discurso del método*, p. 43). Pero en ese reparto las máquinas no están incluidas. Así como durante las décadas de los 50 y los 60 los investigadores de la IA simbólica se centraron en la heurística y el aprendizaje, y en los 70 el tema principal fue la representación del conocimiento, en los 80 la gran preocupación de la IA simbólica fue dotar a las computadoras electrónicas de *sentido común* (*common sense*) (Crevier, 1993, p. 237), una cualidad que, entre otras aplicaciones, habría salvado del colapso a los sistemas expertos. Conseguir que las máquinas fueran capaces de ver, entender y hablar el lenguaje natural se convirtió en una prioridad.

El proyecto estrella, por resultados y por financiación, fue el CYC de Douglas Lenat, del que ya hablamos en el capítulo anterior. Se trata de una IA simbólica iniciada en 1984 que funciona con lógica no monotónica y que se divide en las dos partes del iceberg que antes hemos utilizado como metáfora de la división del conocimiento en consciente e inconsciente: en la parte superior, una enciclopedia, y en la inferior, un conjunto de enunciados de sentido común lo suficientemente grande como para entender las entradas de la enciclopedia. Según Lenat, en 1987 se alcanzó la fase de *convergencia semántica* (*semantic convergence*), en la cual ya era posible definir nuevos conceptos en base a los anteriores, de manera que el CYC podría aprender de manera casi autónoma. Pero la realidad es que el CYC, a fecha de 2013, continúa aún en desarrollo debido a los problemas planteados por la lógica no monotónica y a que todavía requiere de la laboriosa tarea de codificación a mano de enunciados de sentido común. En opinión de Allen Newell, el CYC está condenado al fracaso porque, debido a la naturaleza de sus mecanismos de representación del conocimiento, dos

ingenieros pueden codificar conceptos relacionados de maneras tan distintas que la máquina no sea capaz de captar su relación (Ibíd., p. 243). Sin embargo, a pesar de la enorme envergadura de dificultades como éstas, Lenat no dudó en saltarse los imperativos mertonianos del desinterés y el escepticismo organizado, como es habitual en la IA, y se atrevió a pronosticar que en 2015 «nadie pensaría en comprar una máquina sin sentido común» (Ibíd., p. 242).

Marvin Minsky, por su parte, publicó en 1986 *The society of mind*, un libro de gran éxito comercial en el que intentaba explicar la inteligencia en su totalidad proponiendo un modelo modular de la mente. En el prólogo Minsky decía que: «El sentido común no es una cosa simple. Más bien es una inmensa sociedad de ideas prácticas adquiridas duramente –de multitud de reglas y excepciones, disposiciones y tendencias, equilibrios y comprobaciones aprendidos en la vida» (Minsky, 1985, p. 22). El sentido común surgiría, por tanto, de la interacción de muchos elementos. A esos elementos independientes él los denomina *agentes* (*agents*), cada uno de ellos capaz de realizar una operación mental muy simple. Los agentes se unen a través de *líneas k* o *líneas de conocimiento* (*knowledge lines*) para formar sistemas especializados llamados *servicios* (*services*), capaces de realizar operaciones más complejas, y que a su vez pueden unirse con otros en una jerarquía creciente. En ese esquema hay dos tipos de agentes especiales: supresores y censores. Los *supresores* (*supressors*) se crean por experiencia para evitar que repitamos conductas perjudiciales, y con el tiempo se convierten en *censores* (*censors*) que, directamente, anulan la contemplación de esas conductas como una posibilidad de acción, acelerando por tanto los procesos de pensamiento. Aunque fue, como decimos, un éxito editorial, el libro de Minsky fue duramente criticado por sus colegas por emplear un estilo demasiado metafórico que no ayudaba a esclarecer de manera precisa e inequívoca cómo la conducta compleja surge de la interacción de partes simples. Terry Winograd, el creador de SHRDLU, acusó a Minsky de «hacer un truco de magia cambiando los agentes "tontos" por homúnculos "inteligentes" que se comunican en lenguaje natural» (Crevier, 1993, p. 257).

Un tercer proyecto de IA simbólica iniciado en los 80, y que continúa activo hoy, es *SOAR*, acrónimo de *State, Operator And Result* (estado, operador y resultado), que tiene por objetivo el dar cuenta de la cognición en su totalidad desde el enfoque de la IA humana. SOAR inició su andadura en Carnegie de la mano de Allen Newell y sus alumnos graduados Paul Rosenbloom y John Laird, siento este último el que lo dirige en la actualidad en la Universidad de Michigan. Su arquitectura es la de un sistema de producción con una estructura de control que dispone de un mecanismo de resolución de problemas a través de varios métodos débiles como el análisis medios-fines. SOAR es capaz de aprender nuevas reglas de producción gracias al *chunking* (troceado), una técnica que consiste en «unir nociones existentes en un conjunto (*bundle*) que se convierte en una nueva noción» (Ibíd., p. 261). De esta manera, las soluciones heurísticas a problemas imprevistos son aprendidas por el sistema e incorporadas a su conocimiento en forma de reglas de producción.

El fracaso persistente del cognitivismo

A día de hoy, CYC y SOAR continúan activos, como decimos, pero siguen sin alcanzar las competencias necesarias para ser calificados como inteligencias artificiales en sentido fuerte. De hecho, no existe ninguna IA, ni simbólica ni subsimbólica, capaz de comportarse casi como un ser humano en ninguna de las dos dimensiones del mundo, ni la física ni la social. En cuanto al mundo social, ningún programa ha conseguido todavía ganar alguno de los dos galardones especiales del *premio Loebner*. Este premio desafía cada año, desde 1990, a los mejores programas de comprensión del lenguaje natural a pasar un test de Turing con ligeras modificaciones. Además de premiarse a los mejores de cada edición, hay dos galardones especiales. Uno de ellos, dotado con 25.000 dólares, es para aquel que logre engañar a los jueces a través de una conversación basada únicamente en intercambio de información escrita en lenguaje natural. El otro, de 100.000 dólares, incluye además información visual y auditiva. Nadie ha ganado ninguno de los dos premios especiales.

Por tanto, no existe ninguna IA capaz de pasar el test de Turing. Ciertamente, esto no implica que no exista ninguna IA fuerte pues, como ya se argumentó en el capítulo anterior, un sujeto, como por ejemplo un chimpancé, puede ser inteligente y sin embargo no superar el test de Turing. No obstante, recordemos que nuestro interés se centra en la reproducción de la inteligencia humana, no de la inteligencia en general. El que ninguna computadora electrónica haya pasado el test de Turing de los premios especiales Loebner es un hecho a tener en cuenta, dado que toda inteligencia es susceptible de ser codificada en un sistema de símbolos (Gardner, 1993, p. 39), y el lenguaje natural, por su plasticidad, es capaz de expresar cualquier sistema de símbolos. Si una máquina comprendiese el lenguaje natural, por lo menos al nivel más bajo de los tres distinguidos por Schank, no cabe duda de que sería considerada socialmente como una verdadera IA (Schank, 1986, p. 124). La inteligencia, más allá de los intentos por definirla científicamente, es un atributo social. Un sujeto capaz de comunicarse en lenguaje natural sería calificado de inteligente, sin perjuicio de que la inteligencia pueda ser demostrada de otras formas. Ya justificamos al final del capítulo anterior la importancia especial de la inteligencia lingüística apelando a la codificabilidad en un sistema de símbolos de cualquier inteligencia y la consiguiente posibilidad de expresarlo mediante el lenguaje natural.

Los programas de comprensión del lenguaje natural mejoran cada año, pero todavía son extremadamente débiles. Su avance es similar al de la tortuga de la paradoja de Zenón, que parece estar condenada a no alcanzar jamás a Aquiles. Este hecho es un argumento muy fuerte contra la IA simbólica, en tanto que los defensores de esta corriente suelen sostener que el lenguaje comparte con la mente una serie de características que sólo son producibles desde su enfoque computacional de la mente, y nunca obtenibles por la IA subsimbólica. El filósofo Jerry Fodor y, paradójicamente por su nombre, el psicólogo Zenon Pylyshyn enumeran tres de dichas características: productividad, sistematicidad y coherencia inferencial (Fodor & Pylyshyn, 1988, p. 328). Estas propiedades son supuestamente comunes al lenguaje y al pensamiento debido a que, según afirma Fodor, la mente opera con una especie de *lenguaje del*

pensamiento (*language of thought*) denominado *mentalés*, y a la inversa, el lenguaje natural es una expresión de las oraciones pensadas en mentalés. Que el pensamiento es productivo quiere decir que podemos tener un número ilimitado de pensamientos. Nadie sabe con certeza si el pensamiento humano es productivo, pero muchos filósofos lo consideran una hipótesis plausible (Copeland, 1993, p. 300). Que es sistemático significa que la capacidad de tener ciertos pensamientos está intrínsecamente conectada con la capacidad de tener otros. Así, que alguien pueda pensar "María ama a Juan" implica necesariamente que también pueda pensar "Juan ama a María". Y, por último, la coherencia inferencial es explicada con el ejemplo de que, cuando un sistema posee coherencia inferencial, es imposible que de la proposición molecular $P \wedge Q \wedge R$ infiera $P \wedge Q$ pero no sea capaz de inferir $Q \wedge R$.

Según Fodor y Pylyshyn, estas tres características del lenguaje y del pensamiento, que son productividad, sistematicidad y coherencia inferencial, sólo son explicables desde un enfoque simbólico de la mente. Semejante afirmación, en primer lugar, es falsa, al menos en lo referente a la sistematicidad, pues a finales de los 80 David Chalmers construyó una red de neuronas con arquitectura *RAAM* (*Recursive Auto Associative Memory*, memoria autoasociativa recursiva) capaz de procesar representaciones implícitas con sistematicidad (Clark, 1992, p. 389). Y, en segundo lugar, el enfoque simbólico de la mente y el lenguaje debería ir más allá de la argumentación filosófica y ser demostrado empíricamente mediante la construcción de una IA fuerte capaz de entender el lenguaje natural. La justificación de por qué no se ha logrado, dicen Fodor y Pylyshyn, es que se trata de un problema técnico, es decir, que la culpa es de los ingenieros, bien porque todavía no han descubierto los algoritmos adecuados, o bien porque todavía no han inventado las computadoras electrónicas necesarias (Fodor & Pylyshyn 1988, p. 338). En cualquier caso, el hecho es que el lenguaje natural permanece fuera del alcance de la IA simbólica.

En cuanto a las facultades para habérselas con el mundo físico, suponen un problema formidable para la IA simbólica. Desde el punto de vista de Fodor y de los partidarios de la IA simbólica en general, para ser coherentes, la *habilidad física* (*skill*)

de subir escaleras debería ser efectuada mediante la formulación y manipulación de enunciados en mentalés. En consecuencia, al subir una escalera se supone que en algún rincón oculto de nuestra mente está teniendo lugar un proceso como el descrito por Julio Cortázar en su relato antes mencionado *Instrucciones para subir una escalera*. Es un poco largo, pero es tan esclarecedor que merece la pena transcribirlo íntegro.

«Nadie habrá dejado de observar que con frecuencia el suelo se pliega de manera tal que una parte sube en ángulo recto con el plano del suelo, y luego la parte siguiente se coloca paralela a este plano, para dar paso a una nueva perpendicular, conducta que se repite en espiral o en línea quebrada hasta alturas sumamente variables. Agachándose y poniendo la mano izquierda en una de las partes verticales, y la derecha en la horizontal correspondiente, se está en posesión momentánea de un peldaño o escalón. Cada uno de estos peldaños, formados como se ve por dos elementos, se situó un tanto más arriba y adelante que el anterior, principio que da sentido a la escalera, ya que cualquiera otra combinación producirá formas quizá más bellas o pintorescas, pero incapaces de trasladar de una planta baja a un primer piso. Las escaleras se suben de frente, pues hacia atrás o de costado resultan particularmente incómodas. La actitud natural consiste en mantenerse de pie, los brazos colgando sin esfuerzo, la cabeza erguida aunque no tanto que los ojos dejen de ver los peldaños inmediatamente superiores al que se pisa, y respirando lenta y regularmente. Para subir una escalera se comienza por levantar esa parte del cuerpo situada a la derecha abajo, envuelta casi siempre en cuero o gamuza, y que salvo excepciones cabe exactamente en el escalón. Puesta en el primer peldaño dicha parte, que para abreviar llamaremos pie, se recoge la parte equivalente de la izquierda (también llamada pie, pero que no ha de confundirse con el pie antes citado), y llevándola a la altura del pie, se le hace seguir hasta colocarla en el segundo peldaño, con lo cual en éste descansará el pie, y en el primero descansará el pie. (Los primeros peldaños son siempre los más difíciles, hasta adquirir la coordinación necesaria. La coincidencia de nombre entre el pie y el pie hace difícil la explicación. Cuídese especialmente de no levantar al mismo tiempo el pie y el pie). Llegando en esta forma

al segundo peldaño, basta repetir alternadamente los movimientos hasta encontrarse con el final de la escalera. Se sale de ella fácilmente, con un ligero golpe de talón que la fija en su sitio, del que no se moverá hasta el momento del descenso**»**. Esta narración no es otra cosa que un algoritmo, aunque sea en lenguaje natural, para realizar explícitamente algo que todos hacemos implícitamente: subir escaleras.

John Haugeland enumera tres objeciones contra la posibilidad de codificar algorítmicamente las habilidades físicas en enunciados y de que esa codificación sirva para ejecutarlas tal y como exige la metáfora computacional (Haugeland, 1978, p. 272). Primero, da igual cuán detalladas sean las instrucciones para ejecutar una acción como subir escaleras, porque no son ni *necesarias* ni *suficientes*. Segundo, porque, si bien algunas de ellas pueden ejecutarse de manera consciente e intencional paso por paso, los verdaderos expertos las realizan sin pensar. Ésta es la diferencia señalada anteriormente entre el novato al volante y el conductor veterano. Y tercero, porque las habilidades físicas son más rápidas que el pensamiento. Así, un pianista jamás podría interpretar una obra si tuviera que pensar una por una en cada nota que viene a continuación. Por tanto, el *pensamiento declarativo* o *explícito* no es necesario ni suficiente para dar cuenta del *pensamiento procedimental* o *implícito*.

En este punto los cognitivistas como Fodor tienen dos alternativas: o aferrarse a su teoría del mentalés hasta el final y seguir defendiendo que la cognición se basa en la manipulación de acuerdo a reglas de representaciones cuasi-lingüísticas, o bien renunciar a dar una explicación integral de la cognición. Esta segunda opción es, dice Haugeland, extremadamente peligrosa. Él la denomina *estrategia de segregación*, y señala que su peligro estriba en que, una vez se ha renunciado a explicar las habilidades físicas, se abre la puerta a la posibilidad de que otras habilidades también tengan otro tipo de explicación. Por ejemplo, la percepción, dice Dreyfus, es resultado de un aprendizaje procedimental (Dreyfus, 1992, p. 249). Aprendemos a ver y a oír igual que aprendemos a caminar y a subir escaleras: sin manipular expresiones cuasi-lingüísticas, y sin que haya expresiones de ese tipo que puedan servir para aprender a percibir. En cuanto a la primera opción, la de defender el mentalés hasta el final, es la

de Fodor y Pylyshyn. Ambos reconocen que hay regularidades conductuales que no pueden estar determinadas por reglas implícitas, sino que deben serlo por reglas implícitas. Ahora bien, a su entender la diferencia entre implícito y explícito es un tanto extraña, pues lo implícito, dicen, corresponde al nivel del código máquina, mientras que lo explícito son las instrucciones en lenguaje de programación de alto nivel. Es lo que se desprende de sus palabras: «No todas las funciones de una computadora clásica pueden ser codificadas en la forma de un programa explícito – algunas deben ser cableadas (*wired in*). De hecho, el programa entero puede ser cableado (*hard-wired*) en aquellos casos en los que no necesita modificarse ni examinarse a sí mismo. En tales casos, las máquinas clásicas pueden ser de *reglas implícitas* (*rule implicit*) con respecto a sus programas, y el mecanismo de sus transiciones de estado es enteramente subcomputacional (es decir, subsimbólico). Lo que necesita ser explícito en una máquina clásica no es su programa sino los símbolos que escribe en sus discos (o almacena en sus registros). Éstos, sin embargo, corresponden no a las reglas de transición de estados de la máquina, sino a sus estructuras de datos» (Fodor & Pylyshyn, 1988, p. 342).

Atendiendo a esta peculiar noción de "implícito" y "explícito", para Fodor y Pylyshyn un programa implícito sería el escrito en la tabla de instrucciones de una máquina de Turing o, lo que es igual, de una computadora controlada por programa, mientras que ese mismo programa sería explícito si fuera escrito en la cinta de una máquina universal de Turing o, su equivalente en la realidad, una computadora de programa almacenado. Según ellos, si el programa es inmodificable, es mucho más que implícito: es subcomputacional y subsimbólico. Esta noción de "explicitud", dice el filósofo Andy Clark, es síntoma de una patología a la que él denomina *fijación de código* (*code-fixation*) (Clark, 1992, p. 377). La fijación de código se caracteriza por utilizar un criterio *estructural* de explicitud, según el cual explícita es aquella representación que se da en la superficie de una estructura de datos. Así, una regla escrita en la cinta de una máquina universal de Turing sería explícita en virtud de su accesibilidad para el programador. En cambio, Clark propone un criterio *funcional* de

explicitud, por el que lo implícito y lo explícito serían los dos extremos de un continuo. El grado de explicitud de una información sería mayor cuanto mayor fuese su accesibilidad no para un observador externo, sino para el propio sistema que procesa la información, entendiendo la accesibilidad como una variable bidimensional que depende tanto de la *comodidad* (*ease*) para usar la información como de la *variedad* (*variety*) de sus modos de uso (Ibíd., p. 384).

Estemos o no de acuerdo con la propuesta alternativa de Clark, lo indiscutible es que la noción de explicitud de Fodor y Pylyshyn es un disparate. Es fácil apreciar que, en el fondo, se basa en la distinción entre hardware y software que ya refutamos en el capítulo tercero. Según ellos, si las instrucciones para subir una escalera de Cortázar están cableadas en el hardware entonces son implícitas, pero si por el contrario están escritas al nivel simbólico del software entonces son explícitas; como si las reglas cableadas en el hardware no fueran cadenas de símbolos. Desde luego, esta confusión no puede deberse a falta de conocimientos sobre ingeniería informática, pues Fodor y Pylyshyn han sido un destacados exponentes del cognitivismo, y por tanto es razonable suponerles un saber profundo sobre el tema. La única explicación posible es, por tanto, que se trata de una burda mentira a sabiendas: no tienen escrúpulos, son capaces de escribir cualquier cosa antes que renunciar al cognitivismo, el paradigma que tanta gloria les ha dado.

En su demencial huida hacia adelante, Fodor y Pylyshyn llegan a tal punto de desvergüenza que justifican el dualismo de la metáfora computacional argumentando que la psicología es tan independiente de la neurociencia como la geología lo es respecto de la química (Fodor & Pylyshyn, 1988, p. 346). Cierto es que la *psicología popular* (*folk psychology*), que es la que todo el mundo posee, es independiente de la neurociencia, en tanto que la gente común tiene teorías psicológicas sin saber nada de neurociencia. De modo análogo, la *geología popular* también es independiente de la química, pues por ejemplo los conocimientos geológicos de los agricultores no suelen ir acompañados de conocimientos sobre química, o al menos no sobre química a nivel académico. Ahora bien, la *psicología científica* es tan dependiente de la neurociencia

como la *geología científica* lo es de la química. Abrir un libro cualquiera de geología científica es abrir un libro de química científica. Desde el punto de vista científico los volcanes, las montañas y la deriva continental son fenómenos cuyas propiedades observables a simple vista dependen de las propiedades de las sustancias químicas que participan en ellos, aunque no sean enteramente reducibles a las explicaciones nomológico-deductivas de la química, y requieran también de explicaciones morfológicas y sistemáticas de nivel superior.

Por último, respecto a las escasas habilidades físicas que de hecho son realizables por inteligencias artificiales simbólicas, es notorio que todas padecen el mal de la *limitación de dominio* (*domain limitation*) (Crevier, 1993, p. 250), es decir, que sólo se desenvuelven bien en contextos restringidos. Un ser humano normal tiene una inteligencia espacial y una inteligencia cinético-corporal muy versátiles que le permiten desde botar una pelota hasta pilotar un avión. En cambio, el piloto automático de un avión, como el del bombardero B-24 utilizado por Minsky para confeccionar su primera IA, aunque sea capaz de ejecutar piruetas imposibles para un ser humano, no puede botar una pelota. En última instancia, son las instituciones sociales que conforman lo que Gardner denomina el *ámbito* (Gardner, 1993, p. 64) las que deciden si una conducta es inteligente o no, pero teniendo en cuenta que lo que caracteriza a la inteligencia humana es su capacidad para hacer un intento pasable en casi cualquier cosa, sería extraño que en algún contexto cultural, y particularmente el nuestro, se le concediera el grado de verdadera IA a un piloto automático. Las causas de la limitación de dominio son de tipo social y técnico. De las sociales nos ocuparemos en el capítulo octavo, y de las técnicas, en el siguiente.

7. Condiciones de posibilidad técnicas

Finalizada ya la parte analítica de nuestro estudio, ha llegado el momento de realizar la síntesis. Recapitulemos: en el capítulo tercero describimos las características formales, materiales y pedagógicas de las computadoras electrónicas; en el cuarto, las tesis principales de la psicología cognitiva y de la neurociencia, que son las disciplinas en las que se basan respectivamente la IA simbólica y la IA subsimbólica; en el quinto, tres teorías de la inteligencia: la de Schank desde el enfoque mentalista de la psicología cognitiva, la de Hawkins desde el enfoque fisicalista de la neurociencia, y la de Gardner como teoría de la inteligencia integradora de la mente y el cerebro que suscribimos como la más acertada a nuestro juicio; y en el sexto, hemos repasado la Historia de la IA para observar cuáles han sido sus dificultades técnicas. En el presente capítulo vamos a sintetizar todos estos elementos para elucidar si las inteligencias artificiales en sentido fuerte simbólicas y subsimbólicas son técnicamente posibles por principio, desde el punto de vista de la epistemología.

7.1. Problemas de la IA simbólica

Acabamos de exponer la Historia reciente de la IA, desde que se comercializaron las primeras computadoras electrónicas a mediados del siglo XX hasta la actualidad. No obstante, como ya señalamos al comienzo del capítulo segundo, el anhelo por duplicar las facultades intelectuales del hombre data de muy antiguo. Concretamente, en opinión de Hubert Dreyfus, la IA comienza con Platón en el siglo IV a.C. (Dreyfus, 1992, p. 67), cuando su maestro Sócrates le pide a Eutifrón que defina la piedad: «Ese carácter distintivo es lo que yo quiero que me hagas manifiesto, para que,

considerándolo con atención y sirviéndome de él como de un modelo, pueda declarar que todo lo que tú u otro hace de igual modo es piadoso, en tanto lo que difiere de él no lo es» (Platón, *Eutifrón*, 6e). Ya señalamos en el capítulo quinto que desde Platón en Occidente la sabiduría se ha identificado con la expresión lingüística. El método filosófico de Sócrates, la *mayéutica*, consiste en auxiliar a otros en el parto intelectual para dar a luz definiciones que reflejen la verdad, es decir, las esencias que hacen que las cosas sean como son. Lo que Sócrates pide a Eutifrón es una definición de la esencia de la piedad que sirva como *algoritmo* para que cualquiera que lo aplique sea capaz de distinguir las conductas piadosas.

Sin embargo, dice Dreyfus, Platón no es todavía un cibernético, porque los algoritmos que busca, al estar formulados en un lenguaje natural, presuponen que el sujeto que quiera aplicarlos tenga un conocimiento *semántico*, del significado de los términos empleados para describirlos. Así se observa en el *Menón*, cuando éste replica a Sócrates acerca de la definición de lo que es una figura: «Dices que la figura es lo que siempre acompaña al color. Sea; pero si tu interlocutor declara ignorar lo que es el color y carecer tanto de la experiencia acerca de esta cuestión como acerca de la figura, ¿crees tú que tu definición valdrá algo?» (Platón, *Menón*, 75c). A lo que Sócrates responde: «Por mi parte, la creo verdadera y, si tuviera que vérmelas con uno de estos hábiles o sabios que no buscan más que disputas y querellas, le diría: "Mi respuesta es ésta; si me equivoco, a ti te corresponde hablar y refutarla". Pero, cuando son dos amigos, como tú y yo, los que tienen ganas de charlas, hay que emplear más dulzura y formularlas de una manera más conforme al espíritu de la conversación. Ahora bien: me parece que lo que caracteriza este espíritu no es tan solo responder la verdad, sino también fundamentar la respuesta de uno únicamente en lo que el mismo interlocutor reconozca saber» (Ibíd., 75d).

La última sentencia deja claro que las definiciones se fundamentan en el conocimiento del interlocutor sobre el significado de los términos empleados. Esto es un inconveniente, pues quien escucha una definición puede darle a las palabras de las que se compone significados diferentes a los que tenía en mente quien la formuló.

Sócrates estaba al tanto de esta debilidad de la escritura, y es la razón por la que no quiso dejar testimonio escrito de sus reflexiones. Tal y como dice en el *Fedro*: «Es eso [...] lo terrible que tiene la escritura y que es en verdad igual a lo que ocurre con la pintura. En efecto, los productos de ésta se yerguen como si estuvieran vivos, pero si se les pregunta algo, se callan con gran solemnidad. Lo mismo les pasa a las palabras escritas. Se creería que hablan como si pensaran, pero si se les pregunta con el afán de informarse sobre algo de lo dicho, expresan tan sólo una cosa que es siempre la misma. [...] Y cuando (el escrito) es maltratado, o reprobado injustamente, constantemente necesita de la ayuda de su padre, pues por sí solo no es capaz de defenderse ni de socorrerse a sí mismo» (Platón, *Fedro*, 275d).

La única manera de proteger a un texto de las interpretaciones no deseadas por su autor es renunciando al lenguaje natural y escribiéndolo en un lenguaje formal, debido a que los lenguajes formales, como apuntamos en el capítulo tercero, están exentos de interpretación semántica en su definición (Falguera & Martínez, 1999, p. 61). Ciertamente, pueden ser interpretados (Ibíd., p. 59), refiriendo los símbolos de su vocabulario a objetos del mundo, pero ni es necesario ni afecta a la validez de los enunciados formados con dichos símbolos, porque la validez depende de la forma, no del contenido material. Por ejemplo, dada la expresión matemática x=y+10, el significado de x, que es un valor numérico, variará en función del valor de y, pero siempre de una manera ajena a su significado, es decir, con independencia de que x e y se interpreten como refiriéndose al número de manzanas que hay en una cesta o al número de personas que hay en una habitación.

Para que el proyecto racionalista iniciado por Platón alcanzase su meta, señala Dreyfus, había que eliminar la apelación a la intuición y al juicio, y eso es precisamente lo que hacen los lenguajes formales. Los algoritmos de los programas informáticos, escritos en lenguajes formales, son instrucciones ejecutables de manera rutinaria. Que una instrucción es rutinaria quiere decir, recordemos, que no hace falta ingenio ni perspicacia para llevarla a cabo (Copeland, 2004, p. 43). Por tanto, la definición algorítmica de la piedad habría satisfecho a Sócrates hasta tal punto que, de haberla

hallado, la habría dejado por escrito para la posteridad, pues no habría habido posibilidad de que hubiera sido malinterpretada en tanto que no requeriría interpretación alguna. Los lenguajes formales dicen lo que dicen sin que para entenderlos sea necesario tener contenidos semánticos, o lo que es lo mismo: sin saber nada acerca del mundo. Hasta una máquina de Turing, que jamás ha visto un color o una figura, es capaz de ejecutar algoritmos para manipular colores y figuras, con la sola condición de que hayan sido codificados en un lenguaje formal. Mientras que las palabras enunciadas en un lenguaje natural son, como dice Sócrates, hijas de su padre y necesitan que éste las defienda de las interpretaciones espurias, las palabras de los lenguajes formales no son hijas de nadie, sino que son increadas, eternas, inmutables y universales. Son las Ideas de Platón. Nótese que la comprensión de las Ideas, al igual que la comprensión de las expresiones formales, no requiere del conocimiento previo de ninguna cosa. Muy al contrario, son ellas lo previo en virtud de cuya semejanza comprendemos todo lo demás, según la teoría de la reminiscencia.

La creencia de que la formalización de todo el conocimiento es posible, dice Dreyfus, pronto dominó la filosofía occidental. En el capítulo segundo ya vimos que, en el siglo XVII, Hobbes fue uno de los primeros en expresar explícitamente la concepción del pensamiento como cálculo. Poco después, Leibniz creyó haber encontrado la *mathesis universalis*, el lenguaje universal con el que opera la razón, de tal forma que, dada una controversia cualquiera, sería posible resolverla efectuando las pertinentes operaciones matemáticas (Dreyfus, 1992, p. 69). No obstante, la máquina que diseñó para realizar esos cálculos nunca llegó a funcionar. El proyecto de Leibniz fue continuado en el siglo XIX por otros dos genios. Por un lado, Charles Babbage, quien en 1834 finalizó el primer diseño de su máquina analítica, y por otro, George Boole, que justo veinte años después publicó *The laws of thought*, una obra dedicada a «investigar las leyes fundamentales de aquellas operaciones de la mente mediante las cuales el razonamiento es efectuado, para expresarlas en el lenguaje simbólico de un cálculo» (Ibíd., p. 70). Habría que esperar hasta el siglo XX para la aparición de las computadoras electrónicas, las máquinas perfiladas en todos aquellos trabajos.

El paradigma cognitivista, al basarse en el supuesto nuclear de que la mente es un procesador de información similar a una computadora electrónica, está en continuidad con esta línea filosófica iniciada por Platón. John Haugeland lo expresa así: «Cuando los racionalistas tomaron la cognición como la esencia del ser humano (*res cogitans*), se referían especialmente a la cognición teorética, como la de las matemáticas y la física matemática. La comprensión manifestada en las artes y la artesanía no era, desde su punto de vista, un fenómeno diferente, sino sólo teoría imperfecta, ensuciada por oscuridad y confusión. El cognitivismo es heredero de esta tradición: ser inteligente es ser capaz de manipular (de acuerdo a reglas racionales) representaciones cuasi-lingüísticas de manera "clara y distinta"» (Haugeland, 1978, p. 276). Por su parte, la IA simbólica, al ser una disciplina surgida en el seno del cognitivismo, también pertenece al mismo proyecto filosófico (Dreyfus, 1992, p. xi). Pero no de una manera cualquiera. Lo que la IA tiene de especial y que la diferencia del resto de las ciencias cognitivas es su singular capacidad para demostrar las tesis empíricamente. Mientras que Platón, Hobbes y Leibniz podían plantear sus ideas sobre la mente humana de forma ambigua en proposiciones del lenguaje natural, el investigador de la IA está obligado a detallarlas con la precisión total propia de los lenguajes formales, pues de lo contrario la computadora no las ejecutará.

Por vez primera en la Historia existen unos artefactos, las computadoras electrónicas, que permiten demostrar la tesis racionalista de que el pensamiento consiste en una manipulación de símbolos semejante a la que se realiza para resolver un problema matemático. Ésta es la razón por la cual la computadora no es una analogía más del cerebro o de la mente entre otras muchas surgidas en épocas pasadas, como la de la tabla de cera de Aristóteles, la de la máquina de vapor de Freud, o la del telégrafo de von Helmholtz para el sistema nervioso en general. Por tanto, así como vimos en el capítulo segundo que en la IA fuerte en general se dirime la cuestión *moral* de la muerte del Padre, en la IA fuerte simbólica en particular se decide además una de las grandes cuestiones *epistemológicas* que atraviesan la Historia de la filosofía occidental desde la Antigua Grecia hasta la actualidad.

7.1.1. Procesos cognitivos no replicables

Compartimos la opinión de Hubert Dreyfus de que la IA simbólica presenta los síntomas de lo que Lakatos denomina *programa de investigación degenerado*, es decir, aquella empresa científica que comienza con grandes promesas, ofreciendo un nuevo enfoque que produce resultados valiosos en un cierto dominio, pero que llega un momento en el cual se estanca y deja de progresar (Ibíd., p. ix). A lo largo de esta sección y de la siguiente expondremos los argumentos de Dreyfus contra la posibilidad técnica de la IA simbólica, y por tanto también contra el sueño racionalista iniciado por Platón. Lo haremos siguiendo algunos capítulos de su obra *What computers still can't do*, de 1992. En esta sección repasaremos los cuatro procesos cognitivos de la mente humana que, desde la perspectiva de la psicología fenomenológica de Dreyfus, jamás podrán ser replicados a nivel simbólico por las computadoras. Es, por tanto, una crítica que, en principio, parece afectar sólo a la IA humana. Y en la siguiente descubriremos los cuatro supuestos que subyacen a la IA simbólica: biológico, psicológico, epistemológico y ontológico. Algunos son propios de la IA humana, otros lo son de la IA ajena, y en general son la causa del pertinaz optimismo que empuja a los investigadores de la IA simbólica a creer en la viabilidad de su programa de investigación a pesar de sus síntomas de degeneración.

El ajedrez como campo de batalla

Para identificar los cuatro procesos cognitivos humanos que a su juicio no son replicables por la IA simbólica, Dreyfus comienza comparando las operaciones mentales de un ser humano con las de una computadora durante una partida de ajedrez. A lo largo de la Historia de la IA, el ajedrez ha sido considerado el juego por excelencia (Crevier, 1993, p. 227). Tal predilección se debe, por un lado, a razones contextuales, y por otro, tradicionales. Empezando por las contextuales, hay que tener

en cuenta que las primeras computadoras electrónicas surgieron en la Guerra Fría, cuando los Estados Unidos y la Unión Soviética estaban trabados en una confrontación propagandística a escala global. Para convencer a los neutrales de que se unieran a uno de los dos bloques, la mejor campaña publicitaria era demostrar la superioridad no sólo armamentística, sino también intelectual. Tan importante como fabricar cabezas nucleares era demostrar al mundo que a ese lado del telón de acero estaban las cabezas más inteligentes. El juego escogido para ello fue el ajedrez, representación del enfrentamiento entre dos fuerzas opuestas, la una blanca y la otra negra, que participan movilizando a todos los estamentos de la sociedad.

Durante la Guerra Fría los torneos de ajedrez se convirtieron en un asunto de Estado, sobre todo para la Unión Soviética. Un caso célebre fue el de la final del Campeonato del Mundo de 1978 en Filipinas. Se enfrentaban Anatoli Kárpov, en representación de la URSS, y Víktor Korchnoi, también nacido en Rusia pero que dos años antes había renunciado a la nacionalidad por motivos políticos. Lo que se decidía era, por tanto, la victoria de la URSS frente a sus desertores. La tensión generada por ese simbolismo se reflejó en las partidas. Así, cuando Korchnoi insistió en utilizar su propia silla, Kárpov exigió que fuera desmotada y analizada con rayos X. A mitad de un juego Kárpov comió un yogur que le fue servido por su equipo, y Korchnoi insinuó que el color y el sabor contenían un mensaje en clave. El asesor psicológico de Kárpov, un tal doctor Zhukar, se sentó en primera fila y pasó las horas mirando fijamente a Korchnoi, por lo que éste se quejó alegando que aquel sujeto trataba de hipnotizarlo, y respondió introduciendo en la sala a dos miembros de una secta hinduista. Episodios como éste dejan claro que, en la época en la que nacieron las computadoras electrónicas, el ajedrez era algo más que un juego.

En cuanto a las razones tradicionales por las que la IA ha considerado al ajedrez como su juego predilecto, para encontrar el primer supuesto autómata ajedrecista tenemos que remontarnos hasta finales del siglo XVIII. Concretamente en 1769 el mecánico húngaro Wolfgang von Kempelen inventó una máquina presentada por él como un autómata jugador de ajedrez para entretener a la Emperatriz María Teresa

(Guijarro & González, 2010, p. 327). El artefacto consistía en un maniquí, conocido como "el turco", con un brazo articulado que movía las piezas de un tablero situado sobre un cajón de madera del tamaño de un escritorio. Enseguida empezó a circular la hipótesis, años más tarde comprobada, de que el ingenio era un fraude, y que en el interior del cajón se escondía un ajedrecista humano que movía el brazo del maniquí a través de un pantógrafo. Pero los rumores de descrédito no mermaron la popularidad del aparato. Tras la muerte de Kempelen en 1804, pasó a manos del mecánico vienés Johann Maelzel, quien lo explotó exhibiéndolo en diversas ferias por Europa y América. Babbage lo vio en dos ocasiones y, aunque estaba convencido de que se trataba de un engaño, le sirvió para afianzar su creencia en la posibilidad de crear un autómata capaz de jugar a cualquier juego de habilidad (Ibíd., p. 341).

El primer autómata ajedrecista verdadero fue el construido en 1914 por el ingeniero español Leonardo Torres Quevedo, basándose en un complejo sistema de ruedas dentadas e imanes. Ya en la era digital, el primero en publicar un artículo sobre el ajedrez por computadora fue Claude Shannon en 1950, quien estimó en 10^{120} el número de nodos del árbol de decisión de este juego (Crevier, 1993, p. 223). Poco después, en 1958, vio la luz uno de los primeros programas operativos, el creado por Newell, Simon y Shaw. Sin embargo, era muy torpe, y todavía habría que esperar hasta 1967 para que Richard Greenblatt, un alumno de Minsky en el MIT, presentara la primera IA ajedrecista con un buen nivel de juego. A partir de ahí, las computadoras electrónicas fueron mejorando sin cesar. Como acicate adicional, en 1979 Edward Fredkin, profesor del MIT y hombre adinerado, ofreció tres premios, sin límite de tiempo para lograrlos (Ibíd., p. 227). El primero, de 5.000 dólares, para el primer programa que alcanzase el nivel de maestría en un torneo contra humanos, fue obtenido en 1983 por Belle, obra de Ken Thompson y Joe Condor de Bell Labs. El segundo, de 10.000, para el que alcanzase el nivel de gran maestro internacional, fue entregado en 1988 a Deep Thought, un proyecto iniciado en Carnegie y finalizado por IBM. Y el tercero, de 100.000, para el que lograse derrotar al campeón del mundo, fue conseguido por Deep Blue, de IBM, por vencer a Gary Kaspárov en 1997.

Desde el punto de vista de la razón instrumental y de la IA ajena, poco importa cuáles sean los procesos que utilizan las computadoras para jugar al ajedrez, pues lo único importante es el resultado. Pero a Hubert Dreyfus sí le interesa analizar cómo lo hacen para compararlo con los procesos mentales de los seres humanos y descubrir diferencias insalvables. Adelantamos ya que la productividad de las reflexiones de Dreyfus se debe a que realiza su análisis de los programas de IA desde las categorías de ciertas escuelas de la psicología comprensiva, como la fenomenología y la Gestalt, que, como dijimos en el capítulo cuarto, plantean un enfoque *molar*, que va de lo general a lo particular, *de arriba a abajo* (*top-down*), para dar cuenta de los fenómenos más complejos describiéndolos tal como se presentan y en su relación de circularidad hermenéutica con los fenómenos más simples. Frente a este planteamiento, el de la IA simbólica, por pertenecer al paradigma cognitivista, es justamente el opuesto, pues se basa en un enfoque *molecular*, que va de lo particular a lo general, *de abajo a arriba* (*bottom-up*), para dar cuenta de los fenómenos complejos descomponiéndolos en otros más simples o atómicos (Dreyfus, 1992, p. 211). Así, al examinar de arriba a abajo unos programas que han sido diseñados de abajo a arriba, es como Dreyfus detecta las deficiencias de la IA simbólica con pretensiones de IA humana. Después veremos por qué dichas deficiencias también afectan a la IA ajena.

La periferia de la conciencia

Al explicar las estrategias heurísticas vimos que, debido a la cantidad astronómica de nodos que conforman el árbol de decisión del ajedrez, las inteligencias artificiales juegan evaluando uno por uno los posibles movimientos y respuestas del adversario dentro de una amplitud y profundidad de búsqueda determinados por reglas generales (*rules of thumb*). Herbert Simon, dice Dreyfus, está convencido de que los maestros de ajedrez juegan utilizando ese mismo tipo de heurística. Sin embargo, esta creencia se revela errónea a la luz del hecho comprobado de que un maestro sólo calcula entre 100 y 200 movimientos por turno, un número insuficiente para jugar con

solvencia, pues para jugar a un nivel mínimamente competente, el programa de Greenblatt necesitaba calcular 26.000 (Ibíd., p. 103). Esta diferencia tan abismal sugiere que los humanos juegan empleando mecanismos adicionales que van más allá del conteo o cálculo de alternativas una por una.

La respuesta de Dreyfus apela a lo que William James denomina la *periferia de la conciencia* (*fringes of consciousness*), una región del pensamiento en la que acontecen procesos mentales de los que no somos conscientes. Un ejemplo de fenómeno que tiene lugar en la periferia de la conciencia sería la percepción difusa de las caras de la gente mientras buscamos a un amigo entre la multitud. Ciertamente, vemos todos los rostros, pero nuestra atención no se focaliza en ninguno de ellos, sino que permanece suspendida hasta que reconocemos a nuestro amigo, y entonces se centra en él. De manera análoga, dice Dreyfus, un maestro observa el tablero de ajedrez sin reparar en ninguna pieza en particular, hasta que, de repente, su atención se focaliza en el lugar clave: una pieza valiosa desprotegida, un pasillo para enfilar dos torres, o una ocasión para hacer jaque.

Lo que le permite jugar así es la experiencia. Tras muchas partidas en su haber, el maestro reconoce que la jugada presente se parece a otra, y el análisis que en su día hizo de aquélla, incluso sin recordarla explícitamente, le sirve para identificar cuáles son los movimientos que debe considerar en la actual. Cuando un maestro mira el tablero, no ve un conglomerado de fichas, sino *disposiciones con sentido* (*Gestalten*). Por eso el nivel de maestría es proporcional a la capacidad para reproducir con precisión un tablero tras haberlo visto durante sólo unos pocos segundos: porque ha reconocido un patrón (Ibíd., p. 104). Por el contrario, la máquina opera como los novatos, basando todo su juego en la pura razón, evaluando posibles movimientos uno tras otro. Para jugar como los maestros, señala Dreyfus, «lo que se necesita en general es tener en cuenta la manera en la que el *fondo* (*background*) de la experiencia pasada y el historial del juego presente pueden determinar qué es lo que se muestra como una *figura* (*figure*) y atrae la atención del jugador. Pero esta noción gestáltica de figura y fondo no tiene cabida en la computación explícita paso a paso» (Ibíd., p. 105).

La réplica principal contra este argumento consiste en postular que lo que sucede en la periferia de la conciencia es, en realidad, un conteo inconsciente de posibilidades, una por una, a cuyo término el resultado es trasladado a la conciencia para indicarle cuál es la jugada en la que debe focalizarse para iniciar a partir de ella el conteo consciente. Sin embargo, esta hipótesis crea más problemas de los que soluciona, porque si realmente la periferia de la conciencia fuera capaz de calcular tantos miles de movimientos en tan poco tiempo, entonces cabría preguntarse por qué el maestro no sigue operando de esa manera hasta el final. Dreyfus concluye que «no hay ninguna evidencia, ni conductual ni introspectiva, de que el conteo sea el único tipo de procesamiento de información involucrado en jugar al ajedrez» (Ibíd., p. 106). A su juicio, el ajedrez implica dos tipos de conducta: *apuntar hacia* (*zeroing in*), que acontece en la periferia de la conciencia, y el *conteo uno por uno* (*counting out*), que tiene lugar en la conciencia y se ocupa de evaluar explícitamente las alternativas.

Tolerancia a la ambigüedad

En el lenguaje natural, para elegir una interpretación entre las varias posibles de un enunciado es necesaria una gran cantidad de información que el enunciado por sí solo no proporciona. Continuando con el ejemplo que pusimos al hablar de las redes semánticas de Ross Quillian, para desambiguar el significado de "Estuve esperándote en el banco" hay que saber si la palabra "banco" se refiere a un mueble para sentarse o a una entidad financiera. En algunas ocasiones, otros enunciados del mismo discurso ayudan a tomar la decisión. Así, si la frase formara parte de una conversación entre dos personas que hablan sobre sus ahorros, un ingenio basado en la estrategia de Quillian podría desambiguar con un alto índice de acierto el significado de "banco". Pero hay muchas otras ocasiones en las cuales el resto del contexto lingüístico no proporciona la información suficiente, sino que una parte fundamental depende del contexto en sentido amplio, más allá de lo lingüístico. Por ejemplo, dice Dreyfus, "Permanece cerca de mí" puede significar cualquier cosa, desde "No te alejes más de

un metro", cuando se le dice a un niño en medio de una multitud, hasta "No te alejes más de una milla", cuando un astronauta se lo dice a otro que está dando un paseo lunar. La pragmática, como dijimos en el primer capítulo, es la parte de la semiótica que se ocupa de estudiar la manera en que el contexto afecta al significado de las proferencias lingüísticas.

Es evidente, dice Dreyfus, que la desambiguación acontece en la periferia de la conciencia, pues entendemos el lenguaje sin necesidad de procesarlo de manera consciente. Pueden postularse dos hipótesis sobre lo que allí ocurre (Ibíd., p. 108). La primera, que es la defendida por el cognitivismo y por tanto también por la IA simbólica, es que allí se realiza, en cuestión de milisegundos, una gran cantidad de manipulaciones simbólicas, como las que haría un operario encerrado en una sala china llena de libros sobre semiótica, deduciendo conclusiones a partir de las reglas codificadas en esos libros y de la información contextual entrante para convertir la sentencia inicial, que es equívoca, en otra final que sea unívoca. Esta solución, objeta Dreyfus, es problemática porque, tal y como señala Bar-Hillel, la cantidad de información contextual entrante es infinita, y por tanto su viabilidad depende de la solución de otro problema aún mayor: el de la selección del marco, que es el problema, recordemos, de la circularidad entre la selección de un marco que establezca cuál es la información relevante y la necesidad previa de saber cuál es la información relevante para decidir cuál es el marco adecuado.

La segunda hipótesis, que es la defendida por Dreyfus, es radicalmente opuesta a la primera, ya que se basa en la concepción pragmática del lenguaje del segundo Wittgenstein. Dreyfus cita el siguiente fragmento del filósofo alemán: «Somos incapaces de delimitar claramente los conceptos que usamos; no porque no sepamos su definición real, sino porque no existe tal cosa como la "definición" real. Suponer que *debe* haberla es como suponer que cuando los niños juegan con una pelota están jugando a un juego con reglas estrictas» (Wittgenstein, 1935, p. 25). Así es justo como opera la mente según el cognitivismo y la IA simbólica: manipulando "definiciones" supuestamente reales. Por el contrario, según Dreyfus, los seres humanos reducimos

la ambigüedad sin necesidad de eliminarla por completo mediante una reducción a enunciados formales unívocos, y lo hacemos gracias a que el contexto organiza nuestra percepción, organización que da lugar a la exclusión de algunos significados, al igual que el maestro de ajedrez ni siquiera contempla algunos movimientos. «Nuestro sentido de la situación [...] nos permite excluir la mayoría de las posibilidades sin que lleguen a ser consideradas. Llamaremos "tolerancia a la ambigüedad" a esta habilidad para reducir el espectro de posibles significados mediante la ignorancia de lo que, fuera de contexto, serían ambigüedades» (Dreyfus, 1992, p. 109).

Sostener que la mente opera con definiciones es un absurdo que queda patente en la siguiente definición de "cerdito-hucha" elaborada por Eugene Charniak para comprender la historia infantil del regalo de cumpleaños de la que hemos venido hablando: «Los cerditos-hucha (CH en adelante) los hay de todos los tamaños y formas, aunque la forma preferida es la de cerdito. Generalmente el tamaño oscila entre el del pomo de una puerta y una tartera. Generalmente en los CH se guarda dinero, por lo que cuando un niño necesita dinero a menudo irá a mirar en su CH. Usualmente para conseguir el dinero necesitas sostenerlo y agitarlo (de arriba a abajo). Generalmente ponerlo boca abajo facilita las cosas. Hay técnicas menos conocidas como utilizar un cuchillo para ayudar a sacar el dinero por la ranura. Si, cuando es sacudido, no proviene ningún sonido del interior, eso suele significar que no hay dinero en la hucha. Lo agitas hasta que sale el dinero. Asumimos que después de que el dinero ha salido se lo queda la persona que lo agita, a menos que se diga otra cosa. Si no sale el dinero suficiente, entonces sigues agitándolo hasta que reúnes el dinero que deseas, o bien hasta que deja de provenir sonido del interior. [...] En general, cuanto más pesado es el CH, más dinero hay dentro. Algunos CH tienen tapas que pueden ser quitadas fácilmente para sacar el dinero. A veces es necesario romper el CH para sacar el dinero. El dinero se deposita en la ranura de la hucha, y en ese punto ya no estás sosteniéndolo directamente. El dinero es almacenado en los CH para tenerlo a buen recaudo. A menudo el dinero es guardado ahí durante el proceso de ahorrar para comprar algo que uno desea. Los CH son considerados juguetes, y por

tanto pueden ser poseídos por niños. Esta posesión se extiende al dinero que hay dentro. Así, por ejemplo, está mal visto usar el dinero del CH de otro niño. Además, se puede jugar con un CH de la misma manera que con soldaditos de plomo y similares, moviéndolo alrededor y fingiendo que está vivo y haciendo algo» (Crevier, 1993, p. 113). Y todo esto es sólo un fragmento de la definición de "cerdito-hucha". Pensar que ésta es la forma en la que los seres humanos almacenamos el significado de las palabras es, a efectos de la IA simbólica, un planteamiento irrealizable, tal y como ha demostrado la eterna demora del CYC.

Discriminación de lo esencial y lo inesencial

Imaginemos un tablero de ajedrez, con sus 64 casillas blancas y negras, y un montón de galletas rectangulares, cada una de las cuales ocupa exactamente dos casillas del tablero. Obviamente, el tablero puede cubrirse por completo utilizando 32 galletas. Ahora bien, si cortáramos dos casillas, una de una esquina y otra de la esquina opuesta, la pregunta es: ¿podría cubrirse el tablero utilizando 31 galletas? Éste es el llamado *problema del tablero de ajedrez mutilado*, y hay dos formas de resolverlo. Una sería por la *fuerza bruta* (*brute force*), probando todas las combinaciones posibles por *ensayo y error*. La otra requiere una *representación adecuada del problema*, y es la siguiente: cada galleta cubre necesariamente dos casillas de distinto color, por lo que las 31 galletas cubrirán 31 casillas negras y 31 blancas; sin embargo, al haber cortado esquinas opuestas del tablero, se han quitado dos casillas del mismo color, por lo que habrá más casillas de un color que de otro; por tanto, la solución al acertijo es que no, no es posible cubrir el tablero mutilado con 31 galletas. Se trata de un problema resoluble por fuerza bruta, pues 64 casillas dan lugar a un número de combinaciones de galletas explorable una por una. No obstante, si propusiéramos un tablero de ajedrez imaginario de 4^{2048} casillas, la solución mediante la fuerza bruta ya no sería factible en un tiempo razonable. Por tanto, hay soluciones que sólo pueden ser alcanzadas mediante una representación adecuada.

La habilidad que nos proporciona a los seres humanos una representación adecuada de los problemas es, según Dreyfus, la *perspicacia* (*insight*), una noción tomada del psicólogo de la Gestalt Max Wertheimer (Dreyfus, 1992, p. 114). «En esta operación, uno traspasa la estructura superficial y ve el problema básico –lo que Wertheimer denomina "estructura profunda"–, que permite organizar los pasos necesarios para alcanzar una solución». El matemático George Polya, a quien ya nos hemos referido por ser quien acuñó el término "heurística", distingue dos fases en la solución de problemas. La primera depende de la perspicacia: «Primero, debemos entender el problema. Debemos ver claramente cuáles son los datos, cuáles son las condiciones impuestas, y qué es la cosa desconocida que estamos buscando. Segundo, debemos trazar un plan que nos guíe hacia la solución y conecte los datos con lo desconocido» (Ibíd., p. 116). Esta cita a Polya, dice Dreyfus, está incluida en el libro *Plans and the structure of behavior* de George Miller, Eugene Galanter y Karl Pribram, tres eminentes representantes del cognitivismo que cometen el error de minimizar la importancia de la primera fase al afirmar que: «Obviamente, la segunda es la más crítica. La primera es la que hemos descrito en el capítulo 12 como la construcción de una imagen clara de la situación con el objetivo de establecer un test para la solución del problema; es indispensable, por supuesto, pero en la discusión de problemas bien definidos asumimos que ya ha sido obtenida» (Ibíd., p. 116).

En la línea de Miller y compañía, Newell y Simon infravaloran la dificultad de la primera fase en tanto que están convencidos de que se resuelve de la misma forma que la segunda, es decir, aplicando reglas heurísticas, pero en lugar de sobre el problema actual, sobre la experiencia acumulada en la resolución de otros problemas. Recordemos que la paradoja del *Menón* consiste en cómo es posible que podamos buscar algo, en este caso una representación adecuada para un problema, sin conocerlo previamente. Según Newell y Simon: «La paradoja del *Menón* es resuelta por la observación de que la información puede ser recordada, y también la información nueva puede ser extraída del dominio designado por los símbolos» (Newell & Simon, 1975, p. 65). El problema de este enfoque es que la revisión de la

experiencia anterior no puede ser indiscriminada, sino que debe estar orientada a aquellos datos útiles para representar de manera adecuada el problema actual, pero la orientación sólo puede ser proporcionada precisamente por aquello que se busca: una representación adecuada del problema. De lo contrario, por mucho que se revise la experiencia anterior, jamás se encontrará en ella nada útil.

Al entender la primera fase en los mismos términos que la segunda, dice Dreyfus, lo que Newell y Simon están haciendo es semejante a lo que hacían los astrónomos medievales al añadir más esferas compensatorias al sistema ptolemaico (Dreyfus, 1992, p. 114). En su opinión, hablar de heurística en la primera fase «es completamente equivocado, dado que nadie ha tenido éxito en formular las reglas que guían esta elección preliminar o ni tan siquiera en mostrar que en esta fase, donde la perspicacia es requerida, la gente sigue reglas. Por tanto, no hay ninguna teoría computacional del fundamental primer paso en toda resolución de problemas: la capacidad de distinguir entre lo esencial y lo inesencial» (Ibíd., p. 117).

En el caso del tablero mutilado lo esencial es darse cuenta de que la proporción entre casillas blancas y negras ya no es 1:1, mientras que la proporción de casillas blancas y negras cubiertas por cada galleta sí que se mantiene en 1:1. Si el problema ya es difícil de resolver tal y como lo hemos formulado, piénsese en cuánto aumentaría la dificultad para hallar la representación adecuada si en vez de proponerse sobre un tablero de ajedrez se propusiera sobre un patio de baldosas de, por ejemplo, 12x12, todas del mismo color. Sin duda, la alternancia de color de las casillas es una pista clave, mientras que en otros problemas, también planteados sobre un tablero de ajedrez, el color es irrelevante. A través de este ejemplo se observa claramente que no existen los rasgos esenciales en sí mismos, sino que la relevancia depende, no de los objetos por sí solos, sino de su función dentro del *contexto pragmático* configurado por los intereses del sujeto. Dado que la cantidad de contextos pragmáticos es infinita, la distinción de lo esencial y lo inesencial en cada uno de ellos no puede ser determinada de antemano, de manera similar a como el problema de la cualificación revela que no se pueden listar las condiciones de validez de una regla general.

La capacidad de discriminar entre lo esencial y lo inesencial es necesaria tanto para el aprendizaje como para la resolución de problemas. Cualquier configuración previa *ad hoc* limita el dominio de actuación de una IA. He aquí la causa técnica de la limitación de dominio de los sistemas expertos y de todas las inteligencias artificiales con alguna utilidad que se han construido hasta la fecha: en todas esas máquinas los programadores han discriminado de antemano qué es lo esencial, cuáles son aquellas propiedades sobre las que deben aplicar sus reglas heurísticas en la segunda fase. En el caso del ajedrez, por ejemplo, ninguna IA jugadora contempla la importancia del peso y el tamaño del tablero, factores ambos que, por el contrario, son considerados como esenciales por un sistema experto empaquetador. Un ser humano, en cambio, sabe discriminar, en función del contexto pragmático, cuándo el peso y el tamaño son importantes y cuándo no.

Agrupación perspicaz

Michio Kaku, catedrático de física teórica en la Universidad de Nueva York, opina que el reconocimiento de patrones es, junto con el sentido común, el principal obstáculo para el progreso de la IA (Kaku, 2011, p. 113). La capacidad de los seres humanos para reconocer patrones, dice Dreyfus, se debe a la *agrupación perspicaz* (*perspicuous grouping*) (Dreyfus, 1992, p. 128), una habilidad que resulta de la acción combinada de las tres que hemos visto: periferia de la conciencia, tolerancia a la ambigüedad y discriminación de lo esencial y lo inesencial.

Los primeros intentos para dotar de la capacidad de reconocimiento de patrones a las máquinas se basaban en la *normalización* de la información entrante (Ibíd., p. 120). Un ejemplo lo encontramos en la teoría de la visión antes mencionada de David Marr, en la cual la información visual pasa por tres fases sucesivas: bosquejo primario, bosquejo 2½D y representación 3D. Son como tres cedazos superpuestos cuyo propósito es ir eliminando progresivamente el ruido de fondo y realzar las características esenciales. A diferencia de las teorías como la de Marr, observa Dreyfus,

los seres humanos reconocemos los objetos sin eliminar ni realzar nada: simplemente ignoramos lo inesencial. El siguiente paso evolutivo en las técnicas de reconocimiento consistió en la discriminación de características para compararlas con *listados*, sin necesidad de normalización previa. Así, un determinado objeto sería reconocido como perteneciente a una clase cuando poseyera todos o por lo menos la mayoría de los rasgos definitorios de dicha clase. Se trata, ni más ni menos, que la concepción clásica que tenían los griegos de la categorización del mundo (Gardner, 1985, p. 368). El problema de esta estrategia, dice Dreyfus, es, en primer lugar, que presupone la existencia de ciertos rasgos cruciales, y segundo, aun concediendo que existieran, el reconocimiento consistiría en la comparación de un número de rasgos casi infinito con un número enorme de listados. Suponer que todo eso ocurre en la periferia de la conciencia en un instante es tan absurdo como suponer que el maestro de ajedrez calcula de manera inconsciente decenas de miles de movimientos posibles.

Para evidenciar la falta de verosimilitud de esta hipótesis, Dreyfus describe el caso de los pacientes con agnosia estudiados por Kurt Goldstein y Adhemar Gelb. La *agnosia* es la incapacidad de percibir cierto tipo de información a través de las vías sensoriales normales (Kandel, Schwartz & Jessell, 1995, p. 701). En concreto, los pacientes de Goldstein y Gelb padecían una agnosia visual que les impedía reconocer los objetos de manera espontánea. En su lugar, para reconocer un objeto tenían que elaborar un listado de sus características más salientes y buscar una categoría cuya intensión coincidiese con dichas características. Por ejemplo, al ver un triángulo contaban el número de ángulos y llegaban a la conclusión de que era un triángulo. El proceso era largo, fatigoso y lleno de errores, prueba por tanto de que el reconocimiento que tiene lugar en la periferia de la conciencia no se realiza mediante los mismos procesos que el efectuado conscientemente. «El hecho de que nosotros no necesitamos conceptualizar o tematizar los rasgos comunes a varias instancias del mismo patrón para reconocer dicho patrón es lo que distingue el reconocimiento humano del reconocimiento de las máquinas, el cual sólo ocurre en el nivel explícito conceptual de la pertenencia a una clase» (Dreyfus, 1992, p. 123).

Dreyfus prosigue su argumentación señalando tres tipos de reconocimiento de patrones que son imposibles mediante la comparación de listados de características: el reconocimiento de lo genérico, de la semejanza y de la similaridad. El concepto *genérico* (*generic*) lo toma del fenomenólogo Aron Gurwitsch. Un ejemplo sería el de reconocer un bolígrafo como tal. En principio, dice Gurwitsch, podría reconocerse mediante la comparación de un listado de rasgos. El problema es que la determinación de qué rasgos son significativos depende no sólo del objeto en sí mismo, sino de nuestros intereses hacia él. En este punto Dreyfus recurre a la distinción de Wittgenstein entre *síntoma* (*sympthom*) y *criterio* (*criterion*) (Ibíd., p. 124). Los criterios serían los rasgos esenciales, mientras que los síntomas serían accidentales. La cuestión es que la distinción no es fija, sino que varía en función de nuestros intereses y nuestro conocimiento. Se trata del asunto al que nos referimos a propósito del tablero mutilado. La decisión es siempre *ad hoc*, y requiere por tanto una flexibilidad de la que carece cualquier listado empleado por una computadora. Por ejemplo, dice Wittgenstein, para definir una enfermedad algunos rasgos serían considerados como criterios por unos médicos, mientras que otros los valorarían como meros síntomas (Wittgenstein, 1935, p. 25).

El segundo tipo de reconocimiento no replicable mediante la técnica de comparación de listados es la *semejanza* (*resemblance*), un concepto que refiere al reconocimiento que depende fuertemente del contexto. El ejemplo del pato y el conejo de Wittgenstein en sus *Investigaciones filosóficas* es famoso (Wittgenstein, 1953, p. 447). Se trata de un dibujo que, cuando está rodeado de patos, parece un pato, y cuando está rodeado de conejos, parece un conejo. Los cineastas rusos Lev Kuleshov y Vsevolod Pudovkin hicieron un experimento sobre este fenómeno que demostraba lo que después se conocería como el *efecto Kuleshov*. El experimento consistía en proyectar una imagen del semblante inexpresivo de un hombre, el actor Ivan Mosjoukin, después de haber proyectado: primero, un plato de sopa, segundo, una mujer joven muerta, y tercero, un niño jugando con su oso de peluche. El efecto producido era, por una parte, que el hombre daba la sensación de estar mirando hacia

el plato de sopa, la mujer y el niño, y por otra, que al mirar el plato de sopa su semblante parecía pensativo, al mirar a la mujer, transmitía horror, y al mirar al niño, su boca inexpresiva se transformaba en una sonrisa. Por tanto, la interpretación de una misma imagen puede variar de un extremo al opuesto en función del contexto en el que se presenta. El reconocimiento de características aisladas por comparación con un listado no puede dar cuenta de este fenómeno cognitivo.

Finalmente, el tercer y último tipo de reconocimiento de patrones que depende de la agrupación perspicaz es la *similaridad* (*similarity*), concepto que Dreyfus toma también de Wittgenstein (Dreyfus, 1992, p. 126), y que refiere al reconocimiento de objetos como pertenecientes a una "familia" incluso sin compartir ningún rasgo con los otros miembros de dicha familia (Wittgenstein, 1953, §67). «No importa cuál sea la lista de rasgos inconexos que se construya, porque uno siempre puede inventar un nuevo miembro de la "familia" (*family*) cuyos rasgos sean similares a los de los miembros ya dados pero sin ser *exactamente* similar a ninguno de los rasgos de alguno de ellos, y que, sin embargo, en cierta situación sería reconocido como un miembro del mismo grupo» (Dreyfus, 1992, p. 127). Esta peculiar forma de reconocimiento se basa en la combinación de las tres formas de procesamiento de la información vistas antes: periferia de la conciencia, tolerancia a la ambigüedad y perspicacia.

El cómo y el qué

Antes de exponer los cuatro procesos cognitivos que a juicio de Dreyfus no son replicables por la IA simbólica, habíamos señalado que se trata de observaciones que, en un principio, parecen afectar sólo al enfoque realista de la IA humana, pues un defensor del punto de vista instrumental de la IA ajena bien podría objetar que es posible producir conducta inteligente sin replicar la periferia de la conciencia, la tolerancia a la ambigüedad, la perspicacia y la agrupación perspicaz. Esta objeción, aunque perfectamente válida, nos lleva a examinar un asunto problemático para la razón instrumental: *cómo* funciona algo determina *qué* puede hacer.

Recordemos las palabras de Turing, un tanto jocosas, acerca de la IA humana. Si se dedicaran a construir automóviles, decía, los partidarios de la IA humana intentarían diseñar artefactos que se desplazasen utilizando piernas mecánicas, mientras que los de la IA ajena podrían permitirse explorar otras opciones, como las ruedas (Turing, 1948, p. 420). Cuando se inventó la rueda, se inventó un dispositivo capaz de moverse de una manera sin parangón en la naturaleza. Ningún animal conocido utiliza nada semejante a ruedas para desplazarse. Un automóvil se mueve con ruedas, y se mueve. Sin embargo, conviene reparar en que las ruedas no son aptas para moverse, por ejemplo, por terrenos escarpados, donde las cabras, gracias a sus arcaicas cuatro patas, se manejan con destreza. En DARPA lo saben, y por eso, en contra de la opinión de Turing, han invertido millones de dólares en el desarrollo del *LS3*, acrónimo de *Legged Squad Support System*, una especie de cabra mecánica diseñada para transportar material bélico por terrenos accidentados en los cuales ni las ruedas ni las orugas son eficaces. Las ruedas permiten alcanzar una velocidad superior a las patas sobre terrenos lisos, pero en pendientes muy inclinadas son inoperantes. Por tanto, en este sentido, el *cómo* determina el *qué*. Y, a la inversa, en palabras de Zenon Pylyshyn: «Cuando decimos *qué* hace o puede hacer alguien –ya estamos diciendo algo acerca de *cómo* lo hace» (Pylyshyn, 1974, p. 85). Por ejemplo, decir que sube laderas escarpadas es decir que no lo hace mediante ruedas.

Ahora es el momento de aclarar que el experimento mental del "johnsearlio" que propusimos al final del capítulo quinto implicaba mucho suponer, porque no es plausible que un elemento subatómicamente distinto del oro muestre las mismas propiedades. Cómo es un átomo determina qué propiedades tiene, cómo te mueves determina qué terrenos puedes atravesar, y cómo piensas determina qué puedes pensar y qué conductas puedes producir. Ser capaz de hacer un intento pasable en casi cualquier cosa, que es lo característico de la inteligencia humana, es consecuencia de cómo es nuestra mente. Dado que el cómo determina el qué, la forma ideal de crear una IA fuerte de tipo simbólico pasaría, en primer lugar, por descubrir las leyes del pensamiento, tarea que debería ser realizada por un nuevo paradigma de la psicología

que, superando el problema del método, unificase los enfoques molar y molecular. Sólo así se evitarían críticas como las de Dreyfus, basadas en señalar las deficiencias de un paradigma molecular, como es el cognitivista, desde paradigmas molares, como la fenomenología y la Gestalt.

Sin embargo, aunque esta primera condición ideal para la creación de una IA fuerte simbólica se cumpliera, aún persistiría la imposibilidad del realismo científico y por tanto también de la IA humana, en virtud de los argumentos que señalamos en el capítulo cuarto. No habría certeza de que los enunciados descriptivos de las estructuras y las funciones de la mente hallados son verdaderos, es decir, describen la realidad, ni tan siquiera cuando la conducta producida por ellos en una simulación por computadora fuera indiscernible de la de un ser humano en las mismas circunstancias (Marr, 1977, p. 135), pues el mismo efecto puede resultar de causas distintas. La imperfectibilidad insalvable de la réplica artificial con la que se presenta la IA en el imaginario popular apunta hacia la imposibilidad del realismo científico: siempre hay algo importante que se pasa por alto, que no es capturado por los modelos inevitablemente simplificadores elaborados por el pensamiento.

Siendo así que la IA humana es inviable incluso suponiendo una superación del problema del método como la reclamada implícitamente por Dreyfus, el programa de investigación simbólico sólo puede aspirar, en principio, a construir máquinas pensantes desde el punto de vista instrumental de la IA ajena. Dice Dreyfus que empeñarse en explicar los procesos inconscientes del pensamiento en los mismos términos de los conscientes es semejante a lo que hacían los astrónomos medievales cuando añadían más esferas compensatorias al sistema ptolemaico (Dreyfus, 1992, p. 114). Sin embargo, para la IA ajena da igual que un modelo de la mente consista, metafóricamente, en un sistema geocéntrico de esferas compensatorias o en uno heliocéntrico, porque lo único importante desde un enfoque instrumental es que el resultado de las operaciones internas del modelo, es decir, la conducta, sea *casi* como la de un ser humano, tal y como exige la definición de IA fuerte que propusimos en el primer capítulo. Lo que importa es que funcione, el resultado.

El obstáculo para la IA ajena estriba en que el cómo determina el qué, la estructura determina la función, y por tanto requiere dilucidar si los sistemas formales, que es lo que son en última instancia todas las computadoras electrónicas, son capaces de producir una conducta inteligente similar a la de un ser humano, contando con la ventaja de haber sido liberados de los objetivos del realismo científico. La cuestión es si la conducta inteligente puede ser producida por las máquinas de la reducción total: de lo natural a lo formal, de lo equívoco a lo inequívoco, de la totalidad a las partes. Éste será el tema central de la siguiente sección.

7.1.2. Supuestos subyacentes

Dreyfus identifica cuatro supuestos metateóricos como causas del pertinaz optimismo de los investigadores de la IA simbólica (Ibíd., p. 156). El primero es el *supuesto biológico* (*biological assumption*), según el cual el cerebro es una máquina de estado discreto equivalente a una computadora electrónica. El segundo es el supuesto *psicológico* (*psychological assumption*) que, apoyándose en una concepción dualista de la mente, establece que ésta puede ser considerada como un dispositivo que opera con bits de información de acuerdo a reglas formales, igual que un programa informático. El tercero es el *supuesto epistemológico* (*epistemological assumption*), menos informativo que el anterior pues, sin pronunciarse acerca de cómo opera la mente en realidad, simplemente da por sentado que todo el conocimiento puede ser formalizado y por tanto computado por un ordenador. Finalmente, el cuarto es el *supuesto ontológico* (*ontological assumption*) de que la realidad consiste en un conjunto de hechos independientes entre sí desde un punto de vista lógico, un planteamiento inspirado en el atomismo lógico de Bertrand Russell y del primer Wittgenstein, y que fue el fundamento de la gran corriente positivista de la primera mitad del siglo XX: el *positivismo lógico*, también llamado *empirismo lógico* o *neopositivismo*. Veamos las críticas de Dreyfus contra cada uno de estos cuatro supuestos subyacentes a la IA simbólica.

Supuesto biológico

En el capítulo cuarto, al describir el funcionamiento del sistema nervioso, vimos que las neuronas disparan señales o pulsos eléctricos de amplitud constante, de manera que hay disparo o no lo hay, sin posibilidad de que se produzcan señales intermedias de amplitud variable. Esta dualidad del *todo o nada* (*all or nothing*), dice Dreyfus, fue interpretada en los años 50 como equivalente a la dualidad de 1 y 0 con la que operan las computadoras electrónicas, y todavía persiste en la actualidad como justificación del *supuesto biológico* de que el cerebro es una máquina equivalente a un ordenador. Uno de los padres de la informática, John von Neumann, dice lo siguiente en su libro *The computer and the brain* publicado en 1956: «Puedo volver ahora al carácter digital de este mecanismo. Los pulsos nerviosos pueden claramente ser vistos como marcadores de dos valores, en el sentido discutido anteriormente: la ausencia de pulso representa entonces un valor (digamos, el dígito binario 0), y la presencia representa el otro (digamos, el dígito binario 1)» (von Neumann, 1956, p. 43).

Sin embargo, más adelante en el mismo texto, von Neumann pone en duda el carácter nítidamente binario y digital de las neuronas. Las dudas surgen al examinar las complejidades estructurales y funcionales de las neuronas, tales como el periodo refractario (Ibíd., p. 46), la variedad de morfologías sinápticas (Ibíd., p. 54) y la sumación temporal de potenciales electrotónicos (Ibíd., p. 55): «Por tanto, todas las complejidades referidas hasta ahora pueden ser irrelevantes, pero también puede ser que doten al sistema de un carácter (parcialmente) analógico, o de un carácter "mixto"» (Ibíd., p. 60). Dreyfus cita otra obra de von Neumann en la que éste alcanza las mismas conclusiones: «La evidencia disponible, aunque escasa e insuficiente, tiende a indicar que el sistema nervioso humano utiliza principios y procesos diferentes. Así, los trenes de pulsos parecen transportar significado (*meaning*) mediante ciertos rasgos analógicos (incluidos en la notación del pulso –es decir, parece ser un sistema en parte digital, en parte analógico)» (Dreyfus, 1992, p. 160).

Recordemos lo visto en el capítulo cuarto: lo que varía en función de la magnitud en la que se sobrepasa el umbral de activación de una neurona no es la amplitud del potencial de acción producido, que es siempre constante, sino su frecuencia de disparo. Así, la diferencia entre la percepción del color rojo y el amarillo se traduce fisiológicamente no en diferencia de amplitud de las señales eléctricas, sino en diferencias entre las frecuencias de disparo de dichas señales, y la frecuencia es un concepto temporal propio de los *sistemas analógicos*. Los *sistemas digitales*, en tanto que son de estado discreto, son ajenos al tiempo debido a que funcionan a saltos, de un estado a otro obviando los intermedios (van Gelder, 1996, p. 436). «La diferencia esencial entre el procesamiento digital y analógico de información es que en el procesamiento digital un solo elemento representa un símbolo en un lenguaje descriptivo, es decir, transporta un fragmento (*bit*) específico de información; mientras que en un dispositivo que funcione como una computadora analógica las variables físicas continuas representan la información» (Dreyfus, 1992, p. 161).

Dado que la unidad mínima de información en el sistema nervioso no es el bit representado por el potencial de acción, sino la *tasa de disparo* de bits, el cerebro debería ser considerado como un dispositivo analógico. Un ejemplo de dispositivo analógico sería el *analizador diferencial* (*differential analyzer*) inventado en 1931 por Vannevar Bush, del MIT, que servía para resolver ecuaciones diferenciales. Su principal defecto era el mismo que el del compás geométrico de Galileo y demás artefactos de cálculo similares mencionados en el capítulo segundo, y es que los resultados que arrojaba no eran precisos, pues dependían de características físicas, tales como la inclinación angular de un rayo, que no son configurables con total exactitud. Por supuesto, las computadoras electrónicas pueden ser programadas para reconocer frecuencias de disparo de bits, pero en ese caso estarían siendo utilizadas para simular el comportamiento de un sistema analógico, y el supuesto biológico perdería su sentido, pues ya no se debería afirmar que el cerebro es una máquina semejante a una computadora, sino que es semejante a una máquina analógica simulable por una máquina digital como es una computadora electrónica.

Ahondando en el carácter analógico del sistema nervioso, Dreyfus señala que el diámetro de los axones desempeña una función crucial en el procesamiento de la información actuando como filtro (Ibíd., p. 161). La velocidad a la que circula la corriente por un axón está determinada por la resistencia axial y por la capacitancia por unidad de medida del axón (Kandel, Schwartz & Jessell, 1995, p. 156). Y, a su vez, la resistencia axial depende del diámetro del axón y del grosor de su cubierta de mielina. Cuanto más grueso sea el axón y más gruesa sea su cubierta de mielina, más rápido circularán por él los potenciales de acción. Esto da lugar a que la frecuencia final de trenes de potenciales observada en las terminales presinápticas pueda variar respecto de la frecuencia registrada al principio del recorrido en la zona de activación, pues los trenes de potenciales aumentan o disminuyen su velocidad en función del grosor y de la mielinización del axón que atraviesan.

Supuesto psicológico

Refutado el supuesto biológico, el defensor de la IA simbólica se ve obligado a dar un paso atrás y suponer un dualismo de sustancias para defender la equivalencia funcional entre las computadoras electrónicas y la mente. Esto es lo que establece el *supuesto psicológico*: existe un nivel de procesamiento de la información «en el que la mente utiliza procesos computacionales como comparar, clasificar, buscar en listados y además, para producir la conducta inteligente» (Dreyfus, 1992, p. 163). Este supuesto es característico de los partidarios de la IA simbólica humana, como Newell y Simon, aunque también es en el fondo compartido por los de la IA ajena, como Minsky. Según Dreyfus, se trata de una afirmación dogmática, sin base empírica, que surge a causa de una doble confusión: confundir el término "información" en sus sentidos técnico y vulgar y confundir la computación en general con la computación dirigida por reglas.

En cuanto al concepto de "información" (*information*), Dreyfus señala que en la *teoría de la información* de Claude Shannon y Warren Weaver, que es el fundamento de la ciencia informática, tiene un sentido puramente sintáctico, mientras que en su

sentido vulgar implica contenidos semánticos. En palabras de Shannon: «El problema fundamental de la comunicación es el de reproducir en un punto, bien de manera exacta o bien de manera aproximada, un mensaje seleccionado en otro punto. Frecuentemente estos mensajes tienen *significado* (*meaning*); es decir, que refieren a o están relacionados con algún sistema con ciertas entidades físicas o conceptuales. Estos aspectos semánticos de la comunicación son irrelevantes para el problema de ingeniería» (Ibíd., p. 165). A lo que Weaver añade: «La palabra *información*, en esta teoría, es utilizada en un sentido especial que no debe ser confundido con su uso ordinario. En particular, la *información* no debe ser confundida con el significado» (Ibíd., p. 165). Así, en la teoría de la información es indiferente que 1 KB ($8x10^3$ bits) sea la expresión codificada de una página de texto con sentido o de un montón de caracteres aleatorios, porque el problema para transferir 1 KB de una computadora a otra es, en ambos casos, idéntico desde el punto de vista de la ingeniería.

Es exclusivamente en este sentido no semántico en el que debe entenderse la tesis central de la teoría de la información, según la cual la unidad básica de información es el bit, y toda información puede ser reducida a bits, es decir, átomos binarios (Gardner, 1985, p. 37). En cambio, en su sentido vulgar, el concepto de "información" es semántico, y como señala Searle, la sintaxis no es constitutiva ni suficiente para semántica (Searle, 1990, p. 12). Tal insuficiencia es verdadera por definición, pues la sintaxis trata de la forma de los signos, mientras que la semántica se refiere al contenido de éstos. Dreyfus denuncia que mucha de la literatura sobre IA humana debe su plausibilidad al subrepticio y frecuente intercambio entre los usos vulgar y técnico del término "información" (Dreyfus, 1992, p. 166).

Respecto a la otra confusión, la de la computación con la computación dirigida por reglas, Dreyfus señala que, aun concediendo que la mente fuera un procesador de información en el sentido de Shannon y Weaver, ello no implica que dicho procesamiento deba realizarse necesariamente mediante la aplicación de reglas, es decir, siguiendo un programa, pues bien podría suceder que no hubiese programa alguno, como en los sistemas de memoria. Decir que la mente debe necesariamente

procesar la información siguiendo reglas es, recordemos, lo afirmado por la hipótesis fuerte del sistema de símbolos (HFSS) (Copeland, 1993, p. 273) y por la metáfora computacional, y es tan absurdo como decir que cuando los planetas giran alrededor del Sol es necesario suponer que lo hacen resolviendo ecuaciones diferenciales (Dreyfus, 1992, p. 167). Dentro de la IA simbólica, los defensores de la IA humana mantienen la HFSS, así como los psicólogos cognitivistas mantienen la metáfora computacional, sobre la confusión de estas dos nociones: *descriptible por reglas* (*rule describable*) y *gobernado por reglas* (*rule goberned*) (Franklin, 1995, p. 84). Los partidarios de la IA ajena, por su parte, no tienen necesidad de pronunciarse sobre cómo opera la mente humana en realidad, y por tanto están libres de incurrir en esta confusión. Ellos sólo creen que el movimiento de los planetas es formulable en reglas. Pero, incluso suponiendo la formulabilidad de dichas reglas, de ahí no se deduce su existencia: «Del hecho de que todos los procesos fisicoquímicos continuos involucrados en el "procesamiento de la información" humano puedan ser en principio formalizados y calculados de manera discreta no se concluye necesariamente que algún proceso discreto esté teniendo lugar en realidad» (Dreyfus, 1992, p. 168).

Que las leyes de Kepler funcionen no implica que los planetas se muevan calculando sus trayectorias con ellas. Se mueven por la interacción de varias fuerzas, como quizás sean también fuerzas, dice Dreyfus, las causas de nuestros pensamientos y percepciones: «"campos", "fuerza", "configuraciones"» (Ibíd., p. 166), conceptos tomados de la psicología comprensiva. La hipótesis de Dreyfus está avalada por casos como el de Daniel Tammet, un deficiente genial con autismo que es conocido mundialmente por varias habilidades intelectuales asombrosas, entre las que destaca su plusmarca de cálculo de más de 20.000 decimales del número π en 5 horas. Lo fascinante del caso es que Tammet confiesa no calcular las cifras realizando operaciones aritméticas, sino que en su mente ve una sucesión de formas, colores y texturas que le van indicando cuál es el siguiente número (Tammet, 2006, p. 4). Su testimonio es un indicio de que, aunque las inteligencias, como dice Gardner, sean codificables en sistemas simbólicos (Gardner, 1993, p. 39), eso no implica que el

pensamiento opere mediante dichos sistemas. El relato de Tammet se asemeja más a los conceptos de campos, fuerza y configuraciones referidos por Dreyfus. La razón por la que este deficiente genial percibe visualmente esos procesos inconscientes es porque padece sinestesia, una enfermedad de la que ya hablamos en el capítulo quinto, y que consiste en la interferencia de modalidades sensoriales.

La hipótesis del procesamiento de información como causa del pensamiento, concede Dreyfus, es válida como *hipótesis de trabajo*, pero no como *teoría*, que es como la presentan sus defensores. La diferencia entre una hipótesis de trabajo y una teoría es que la primera no necesita explicar todos o por lo menos la mayoría de los fenómenos de los que debería dar cuenta, mientras que la segunda sí (Dreyfus, 1992, p. 171). La realidad es que el supuesto psicológico no es capaz de explicar fenómenos centrales de la inteligencia tales como el uso del lenguaje natural y el reconocimiento de patrones. Por tanto, sólo debe concedérsele el rango de hipótesis. Los investigadores de la IA simbólica cometen el error de basarse en sus exiguos éxitos parciales en la replicación de tareas pueriles como jugar al ajedrez para concluir que el resto de procesos psíquicos han de ser replicables de la misma forma: manipulando símbolos mediante reglas. Su terco convencimiento, contrario a la evidencia, se debe a que manejan una noción positivista de lo que es "explicar", la cual implica la consideración del supuesto psicológico como un axioma verdadero *a priori*.

El psicólogo cognitivista George Miller y sus colaboradores, dice Dreyfus, asumen que «la explicación o descripción completa de una conducta *requiere* describirla en términos de un conjunto de instrucciones» (Ibíd., p. 175). Siendo así que toda explicación se identifica con la reducción a instrucciones o leyes semejantes a las de la física, se sigue necesariamente que la explicación del pensamiento humano debe ser expresable en forma de programa informático. Tal identificación del explicar con las explicaciones de tipo nomológico-deductivo, como si las explicaciones de otros tipos no fueran satisfactorias para la ciencia, es manifiesta en el fundador del positivismo, Augusto Comte, pero se remonta a mucho antes del siglo XIX. Según Dreyfus, su origen está en Platón. Concretamente en el *Menón* el filósofo griego «no

deja duda sobre esto: toda acción sensata (*sensible*), es decir, no arbitraria, tiene una estructura racional que puede ser expresada en términos de alguna teoría, y cualquier persona que realice dicha acción estará siguiendo, cuando menos implícitamente, la teoría tomada como un conjunto de reglas. Para Platón, estas instrucciones ya están en la mente, preprogramadas en una vida anterior, y pueden ser explicitadas planteándoles a los sujetos las cuestiones apropiadas» (Ibíd., p. 176). Volvemos así al asunto de la continuidad del proyecto racionalista en un arco que abarca, entre otros muchos, a Platón, Descartes y Comte, y llega hasta nuestros días en forma de cognitivismo y cibernética; entendiendo el racionalismo en el sentido amplio antes señalado por Haugeland de la creencia en la capacidad de la razón para explicarlo todo, incluida la mente, reduciéndolo a reglas, leyes, instrucciones, algoritmos (Haugeland, 1978, p. 276). Este supuesto apriórico racionalista unido a la confusión entre las nociones de lo descriptible por reglas y lo gobernado por reglas desemboca en la creencia errónea del supuesto psicológico como axioma.

El último argumento de Dreyfus contra el supuesto psicológico apunta a otra confusión implícita en el cognitivismo: la confusión entre el cerebro y la mente y sus respectivos enfoques molecular y molar. Entre el cerebro y la mente, dice Dreyfus, los cognitivistas postulan la existencia de un *tercer nivel* compuesto por fenómenos mentales supuestamente explicables en términos moleculares (Dreyfus, 1992, p. 179). Ulric Neisser, uno de los padres del paradigma cognitivista, pretende explicar así la percepción de la página de un libro: «Si vemos los objetos en movimiento como cosas unificadas, debe ser porque la percepción resulta de un proceso integrativo a través del tiempo. El mismo proceso es sin duda responsable de la construcción de objetos visuales a partir de las sucesivas "instantáneas" (*snapshots*) capturadas por el ojo en movimiento» (Ibíd., p. 182). Dreyfus se pregunta qué son esas "instantáneas". Si son patrones de energía, entonces no puede decirse que sean percibidas a nivel mental, pues los patrones de energía son procesados por el cerebro a *nivel fisiológico*. Pero, por otra parte, tampoco pueden ser fenómenos mentales, pues a *nivel fenomenológico* no percibimos las distintas capturas de una página, sino la página directamente.

Dreyfus acusa a Neisser en particular, y a los cognitivistas en general, de intercambiar la noción de "entrada" (*input*) en sus sentidos físico y perceptual, confundiendo así dos niveles, el cerebral y el mental (Ibíd., p. 181). En esa tierra de nadie (*no-man's-land*), designada por el término "cognición", es donde los psicólogos cognitivistas y los investigadores de la IA humana dan rienda suelta a sus pretensiones positivistas, y se lanzan a suponer que existen procesos mentales de manipulación de símbolos semejantes a los de las computadoras electrónicas, y por tanto descriptibles molecularmente, con el afán último de convertir a la psicología en una ciencia exacta como la física. Es en ese tercer nivel intermedio, oculto, donde se supone que el maestro de ajedrez evalúa miles de jugadas por segundo. Y es ahí donde se supone que residen las estructuras memorísticas de las que habla Roger Schank, como los MOPs, TOPs y scriptlets. Recordemos que el psicólogo norteamericano se mostraba incapaz de resolver el problema de la selección del marco a nivel simbólico (Ibíd., p. 45). Tal vez sea incluso imposible porque, como dice Dreyfus, quizás no exista ninguna estructura simbólica como los marcos. En cambio, a nivel subsimbólico o cerebral vimos en el texto de Jeff Hawkins que el problema quedaba resuelto de manera sencilla mediante el cruce de las conexiones ascendentes de anteroalimentación con las descendentes de realimentación. Así pues, el problema de la selección del marco no se *resuelve* en términos simbólicos, sino que se *disuelve* tan pronto como reparamos en que deriva de aceptar la validez de una hipótesis tan poco plausible como es la del supuesto psicológico (Churchland, 1989, p. 277).

Desde luego, para quien quiera hacerse cargo de toda su problemática, el supuesto psicológico es asumible como hipótesis de trabajo, pero no como axioma universal *a priori*. «El único argumento legítimo para el supuesto de que la mente funciona como una computadora es el de la existencia actual o posible de una máquina inteligente de ese tipo» (Dreyfus, 1992, p. 187). Los psicólogos cognitivistas como Miller, aclara Dreyfus, creen disponer de esa evidencia empírica gracias a los programas de IA humana simbólica como los de Newell y Simon, pero en realidad dichos programas no aportan ninguna evidencia, pues su éxito está restringido a

ciertas tareas, y además se debe a que los programadores han decidido *ad hoc* de antemano qué datos son relevantes, obviando por tanto la primera y fundamental fase en la solución de problemas de las dos distinguidas por George Polya: la consistente en entender el problema, en elaborar una representación discriminatoria de lo esencial y lo inesencial. En definitiva, el cognitivismo y la IA humana simbólica se mueven en la confusión, la oscuridad y la ignorancia de los problemas más difíciles.

Supuesto epistemológico

Sea lo que sea la mente, las objeciones de Dreyfus demuestran que no es obvio que funcione como una computadora electrónica. Los investigadores de la IA simbólica que conceden este punto se ven obligados a dar otro paso atrás y alinearse con Minsky en la defensa de la IA ajena, definida por la hipótesis del sistema de símbolos (HSS). La HSS, recodemos, establece que: «Un sistema de símbolos físico tiene las capacidades necesarias y suficientes para la acción inteligente general» (Newell & Simon, 1975, p. 41). En esto consiste el *supuesto epistemológico*. No dice nada acerca de cómo funciona la mente, sino que se limita a afirmar que la conducta inteligente producida por la mente humana puede ser producida también por un conjunto de algoritmos ejecutados por un ordenador. Por tanto, al ser menos informativo que el supuesto psicológico, es menos falsable, pero aun así es vulnerable. Dreyfus señala dos tesis implícitas en él que pueden ser refutadas, a saber: que toda conducta inteligente (*nonarbitrary*) puede ser formalizada y que dicho formalismo puede ser utilizado para reproducir la conducta en cuestión (Dreyfus, 1992, p. 190).

La creencia de que toda conducta inteligente puede ser formalizada, dice Dreyfus, se basa en el optimismo producido por el éxito en la formalización de dos fenómenos: la física y la sintaxis. Lo que el filósofo norteamericano se propone demostrar es que ambas formalizaciones son de fenómenos cualitativamente muy distintos de la mente, y por tanto la formalización de ésta no está garantizada. Empezando por la física, Dreyfus concede que el organismo humano, en tanto que está

formado en su totalidad por materia sin dejar lugar a una supuesta sustancia espiritual, es formalizable mediante las leyes de la física y la química. En este sentido, la conducta inteligente es, en principio, perfectamente computable. Ahora bien, una IA de este tipo no sería simbólica, sino subsimbólica, dado que no procesaría la información procesada a nivel mental por el cerebro, sino que procesaría un modelo del cerebro. «Una computadora digital resolviendo las ecuaciones que describen un dispositivo analógico de procesamiento de la información y por tanto simulando su *función* no está por ello simulando su "procesamiento de la información". No está procesando la información que es procesada por el dispositivo analógico simulado, sino una *información completamente diferente* referida a las propiedades físicas o químicas del dispositivo analógico. Por tanto, la afirmación fuerte de que *cualquier forma de información* puede ser procesada por una computadora digital es engañosa» (Ibíd., p. 195). Así pues, el éxito de la física no es indicio de que la formalización sea aplicable a la mente, sino, como mucho, al cerebro.

En cuanto a la sintaxis, Dreyfus comienza por señalar el enorme impacto que causó en los años 50 la aparición de la *gramática generativa* (*generative grammar*) de Noam Chomsky. Se trata de una teoría lingüística que permite formalizar cualquier oración reduciéndola a *núcleos oracionales* u *oraciones básicas* (*kernel sentences*) (Gardner, 1985, p. 209). Sobre esos núcleos se aplica un conjunto algorítmico de procedimientos para transformarlos en diferentes cadenas lingüísticas. La utilidad de todo el proceso es la discriminación entre enunciados sintácticamente correctos e incorrectos. Las transformaciones, dice Gardner, ponen al descubierto la ambigüedad de la frase "*The shooting of the hunters disturbed me*" (El tiroteo de los cazadores me perturbó), pues hay dos historias transformacionales que pueden generarla: en una se parte de un núcleo en el que los cazadores disparan, y en la otra se parte de un núcleo en el que los cazadores reciben los disparos. Gardner resume así las pretensiones de Chomsky: «Su idea de la "generación gramatical" se basaba en la concepción de un autómata, una máquina en sentido abstracto, que genera simplemente cadenas lingüísticas basándose en reglas que le han sido incorporadas (programadas). La

gramática resultante es neutral, igualmente válida como descripción de la producción o de la comprensión lingüística. A todas luces, Chomsky era un vástago de la nueva era inaugurada por Wiener, von Neumann, Turing y Shannon» (Ibíd., p. 211).

La gran virtud y al mismo tiempo la gran deficiencia de la teoría de Chomsky, advierte Dreyfus, es justamente el autonomismo de la *sintaxis* (Dreyfus, 1992, p. 198). Los análisis transformacionales se efectúan a partir de criterios puramente sintácticos que ignoran las reglas de la *pragmática*, es decir, del uso de las proferencias lingüísticas. Por tanto, para satisfacer los objetivos del supuesto epistemológico aún haría falta, además de una teoría formal sobre la corrección sintáctica como la de Chomsky, otra teoría también formal sobre la corrección pragmática. «Para que las máquinas se comuniquen en el lenguaje natural sus programas no sólo deben contener las reglas de la gramática; también deben contener las reglas de la práctica lingüística (*linguistic performance*)» (Ibíd., p. 198). En la terminología del segundo Wittgenstein diríamos que la gramática de Chomsky es sólo *de superficie*, y todavía falta una gramática *de profundidad*: «En el uso de una palabra se podría distinguir una "gramática superficial" de una "gramática profunda". Lo que se nos impone de manera inmediata en el uso de una palabra es su modo de uso en la *construcción de la proposición*, la parte de su uso –podría decirse– que se puede percibir con el oído. –Y ahora compárese la gramática profunda de las palabras "querer decir", por ejemplo, con lo que su gramática superficial nos haría suponer. No es de extrañar que nos sea difícil orientarnos» (Wittgenstein, 1953, §664).

Contra la posibilidad de formalizar la pragmática Dreyfus enumera dos argumentos (Dreyfus, 1992, p. 198): un *argumento de principio*, que abordaremos en el siguiente apartado sobre el supuesto ontológico, y que señala que una teoría formal de la pragmática consistiría en una teoría formal de todo el conocimiento humano, lo cual es imposible; y una *descripción objetiva*, de la que nos vamos a ocupar aquí, y que evidencia que no todas las conductas lingüísticas están regidas por reglas. Supongamos la frase "La idea está en el bolígrafo". Para que una máquina entendiese su significado necesitaría una regla para interpretar que, en este caso, "estar en" no tiene un sentido

literal, sino figurado. Sin embargo, los seres humanos somos capaces de entender inmediatamente su significado, aún siendo extraño (*odd*), sin necesidad de ninguna regla. Los hablantes nativos, apunta Dreyfus, reconocemos la extrañeza al tiempo que la superamos sin dificultad. Somos capaces de entender proferencias lingüísticas que contienen «errores gramaticales o semánticos» (Ibíd., p. 199). Prueba de ello son los dos siguientes textos de Julio Cortázar.

El primero, correspondiente al capítulo 68 de *Rayuela*, se caracteriza por una continua violación de las reglas de la *semántica*, pues está repleto de palabras inventadas: «Apenas él le amalaba el noema, a ella se le agolpaba el clémiso y caían en hidromurias, en salvajes ambonios, en sústalos exasperantes. Cada vez que él procuraba reclamar las incopelusas, se enredaba en un grimado quejumbroso y tenía que envulsionarse de cara al nóvalo, sintiendo cómo poco a poco las arnillas se espejunaban, se iban apeltronando, reduplimiendo, hasta quedar tendido como el trimalciato de ergomanina al que se le han dejado caer unas fílulas de cariaconcia. Y sin embargo era apenas el principio, porque en un momento dado ella se tordulaba los hurgalios, consintiendo en que él aproximara suavemente sus orfelunios. Apenas se entreplumaban, algo como un ulucordio los encrestoriaba, los extrayuxtaba y paramovía, de pronto era el clinón, la esterfurosa convulcante de las mátricas, la jadehollante embocapluvia del orgumio, los esproemios del merpasmo en una sobrehumítica agopausa. ¡Evohé! ¡Evohé! Volposados en la cresta del murelio, se sentían balparamar, perlinos y márulos. Temblaba el troc, se vencían las marioplumas, y todo se resolviraba en un profundo pínice, en niolamas de arguntendidas gasas, en carinias casi crueles que los ordopenaban hasta el límite de las gunfias».

El segundo texto pertenece a *La vuelta al día en ochenta mundos*, y destaca por su persistente incumplimiento de las reglas de la *sintaxis*: «Con lo que pasa es nosotras exaltante. Rápidamente del posesionadas mundo estamos hurra. Era un inofensivo aparentemente cohete lanzado Cañaveral americanos Cabo por los desde. Razones se desconocidas por órbita de la desvió, y probablemente algo al rozar invisible la tierra devolvió a. Cresta nos cayó en la paf, y mutación golpe entramos de. Rápidamente la

multiplicar aprendiendo de tabla estamos, datadas muy literatura para la somos de historia, química menos un poco, desastre ahora hasta deportes, no importa pero: de será gallinas cosmos el, carajo qué». Aun violando las reglas de la semántica el uno y las de la sintaxis el otro, estos dos textos de Cortázar son indudablemente significativos, pues sugieren algo al lector.

Si postulásemos que los hablantes disponemos de *metarreglas* (*metarules*) que nos permiten entender las proferencias lingüísticas que, como estos dos relatos, violan las reglas de la gramática, entonces no se entendería por qué, a pesar de poseerlas, sentimos extrañeza ante los enunciados extraños. Y lo que es más importante: se produciría un *regreso al infinito* como el antes observado a propósito del problema de la cualificación, pues siempre serían necesarias reglas de nivel superior para explicar el significado de las proferencias que violan las reglas del nivel actual (Ibíd., p. 200). Para evitar el regreso al infinito es necesario que en algún momento se produzca un corte en la jerarquía de reglas.

La solución del segundo Wittgenstein consiste en afirmar, desde un punto de vista pragmático, que el significado de una palabra es, en una gran clase de casos, su uso en el lenguaje (Wittgenstein, 1953, §43), y ese uso no depende de reglas fijas, sino de convenciones flexibles no explicitadas que varían de un contexto cultural a otro y a través del tiempo dentro de la misma cultura. «En general no usamos el lenguaje de acuerdo a reglas estrictas –ni tampoco lo hemos aprendido mediante reglas estrictas. [...] En la práctica rara vez usamos el lenguaje como un cálculo. No sólo no pensamos en las reglas de uso –definiciones, etc.– mientras usamos el lenguaje, sino que cuando se nos pide que formulemos dichas reglas, en la mayoría de los casos no somos capaces de hacerlo» (Wittgenstein, 1935, p. 25). Aunque los planetas no se muevan resolviendo ecuaciones diferenciales, los físicos pueden describir sus movimientos mediante dichas ecuaciones gracias a que son fenómenos descontextualizados, en el sentido de aislables. Sin embargo, el lenguaje es un fenómeno totalmente *contextualizado*. El significado de las palabras depende de su uso dentro de un *juego de lenguaje* (*Sprachspiel*), y un juego de lenguaje, tal y como lo define Wittgenstein en

las *Investigaciones filosóficas*, no refiere sólo a fenómenos lingüísticos, sino a toda la práctica vital en la que acontece el lenguaje: «Llamaré también "juego de lenguaje" al todo formado por el lenguaje y las acciones con las que está entretejido» (Wittgenstein, 1953, §7). «La expresión "*juego* de lenguaje" debe poner de relieve aquí que *hablar* el lenguaje forma parte de una actividad o de una forma de vida» (Ibíd., §23). Por tanto, formalizar la pragmática implicaría formalizar toda una forma de vida, una tarea que, como vimos en el capítulo cuarto, le resultó imposible al antropólogo Paul Kay (Gardner, 1985, p. 277).

Por su parte, la manera en que los investigadores de la IA simbólica evitan el regreso al infinito está tomada también de Wittgenstein: pero del primer Wittgenstein, el del *Tractatus*. «Para la gente de las computadoras el regreso también se detiene con una interpretación que es auto-evidente, pero esta interpretación no tiene nada que ver con las demandas de la situación. No puede, porque la computadora no está en situación. No genera ningún contexto local. La solución de los teóricos de la computación es construir la máquina para que responda a bits últimos descontextualizados, datos completamente determinados que no requieran interpretación ulterior para ser entendidos» (Dreyfus, 1992, p. 204). Esto es lo que sostiene el *supuesto ontológico*, fundamento de la teoría figurativa del *Tractatus*: que el lenguaje, el pensamiento y el mundo están compuestos de átomos independientes y comprensibles por sí mismos sin necesidad de apelar a ninguna regla o conjunto de reglas en forma de marco o contexto. Así desaparecerían los problemas del regreso al infinito y de la circularidad hermenéutica, que son, a juicio de Dreyfus, los principales obstáculos que han impedido el avance de la IA simbólica (Dreyfus, 1992, p. 60).

Supuesto ontológico

Al comienzo del capítulo anterior vimos que para habérselas con el mundo una computadora puede ser programada mediante *métodos débiles*, basados, como el GPS de Newell y Simon, en la confianza de que la razón pura es capaz de solucionar

cualquier problema, o bien mediante *métodos fuertes*, que además de un sistema de heurística incorporan conocimientos sobre el mundo. Tras el fracaso de los programas como el GPS, en la actualidad nadie duda ya de que los métodos fuertes son el único camino (Ibíd., p. 208). El enfoque de los métodos fuertes, unido al hecho de que las computadoras electrónicas sólo pueden operar sobre elementos determinados e independientes entre sí, que en última instancia son los bits compilados por los lenguajes de programación de alto nivel, arroja la consecuencia de que la IA simbólica sólo es viable suponiendo que el mundo es expresable en forma de una gran masa de hechos discretos. Esto es lo que sostiene el *supuesto ontológico*. Lejos de ser una novedad surgida en la era de la informática, dice Dreyfus, se trata de una creencia profundamente arraigada en la filosofía occidental que tiene su origen en Platón, que fue defendida tanto por los racionalistas como por los empiristas en la Modernidad, y que en el siglo XX converge en la teoría del atomismo lógico propuesta por Bertrand Russell en *The philosophy of logical atomism* y alcanza su máxima expresión en 1921 con la publicación del *Tractatus logico-philosophicus* de su discípulo Ludwig Wittgenstein (Ibíd., p. 211). Repasemos brevemente las tesis principales del atomismo lógico del filósofo alemán.

Las primeras proposiciones del *Tractatus* las dedica Wittgenstein a exponer su visión del mundo. El *mundo* (*Welt*), dice, se compone de *hechos* (*Tatsachen*) (Wittgenstein, 1921, §1.2) que se descomponen en hechos más simples denominados *hechos atómicos* (*Sachverhalten*). A su vez, los hechos atómicos consisten en *objetos* o *cosas* (*Sachen*) simples, irreductibles (Ibíd., §2.02), conectados entre sí de una cierta manera que depende de sus propiedades internas (Ibíd., §2.01231). Esa cierta manera en que se interrelacionan es la *estructura* (*Struktur*) del hecho atómico (Ibíd., §2.032). La totalidad de los hechos atómicos que se dan efectivamente es el mundo (Ibíd., §2.04), mientras que los que no se dan permanecen como meras posibilidades combinatorias. A partir de la proposición 2.1 comienza la exposición de la *teoría figurativa*. Nos hacemos *figuras* (*Bilder*) de los hechos, dice Wittgenstein. Las figuras son modelos de la realidad (Ibíd., §2.12) que pueden representarla gracias a que

guardan con ella una relación de *isomorfismo*: «Lo que cualquier figura, sea cual fuere su forma, ha de tener en común con la realidad para poder siquiera –correcta o falsamente– figurarla, es la forma lógica, esto es, la forma de la realidad» (Ibíd., §2.18). La figuración será correcta, o lo que es igual, verdadera, si el *sentido* (*Sinn*) de la figura, que es lo representado por ella (Ibíd., §2.221), concuerda con la realidad; de lo contrario, será incorrecta, falsa (Ibíd., §2.222). Se trata, por tanto, de la teoría clásica de la verdad como *adequatio rei et intellectus*.

En cuanto al lenguaje y al pensamiento, su función primordial es la figuración. Ambos están íntimamente ligados, como se desprende de las siguientes afirmaciones: «La figura lógica de los hechos es el pensamiento» (Ibíd., §3). «El signo proposicional usado, pensado, es el pensamiento» (Ibíd., §3.5). «El pensamiento es la proposición con sentido» (Ibíd., §4). «La proposición es una figura de la realidad. La proposición es un modelo de la realidad tal como nos la pensamos» (Ibíd., §4.01). En estricta correspondencia con lo dicho acerca del mundo, los hechos, tanto efectivos como posibles, son expresables en lo que Russell denomina *proposiciones moleculares* (*molecular propositions*) (Russell, 1918, p. 37), a las que Wittgenstein se refiere simplemente como *proposiciones* (*Sätze*). A su vez, las proposiciones moleculares son descomponibles en *proposiciones atómicas* (*atomic propositions*) (Ibíd., p. 27), llamadas *proposiciones elementales* (*Elementarsätze*) o *completamente analizadas* por el filósofo alemán, que son aquellas que figuran hechos atómicos. Y, por último, las proposiciones atómicas se componen de *signos simples* o *nombres* cuyo significado son los objetos, que son los constituyentes de los hechos atómicos (Wittgenstein, 1921, §3.202 y §3.203). Las proposiciones atómicas son independientes entre sí hasta el punto de que: «De una proposición elemental no puede inferirse ninguna otra» (Ibíd., §5.134), y su valor de verdad es comprobable empíricamente de manera directa y unidireccional, sin necesidad de conocer otras proposiciones (Ibíd., §2.23).

La consecuencia práctica del atomismo lógico es que el mundo en su totalidad puede ser descrito mediante oraciones atómicas: «La especificación de todas las proposiciones elementales verdaderas describe el mundo completamente» (Ibíd.,

§4.26). En esto consiste el supuesto ontológico. Sin embargo, como señala Dreyfus, se trata tan sólo de una hipótesis, y no de un axioma *a priori*, que es como pretenden presentarlo los defensores de la IA simbólica (Dreyfus, 1992, p. 225). Creer que los seres humanos poseemos los conocimientos sobre el mundo en forma proposicional (supuesto psicológico de la IA humana), como en el ejemplo de Charniak del cerdito-hucha, o que, aun sin tener esa forma, somos capaces de describirlos en ella (supuesto epistemológico de la IA ajena), que es la posibilidad examinada por Cortázar en *Instrucciones para subir una escalera*, son hipótesis que generan problemas de gran envergadura, como por ejemplo el problema del conocimiento. En el capítulo quinto vimos que el *problema del conocimiento* se desglosa en el triple problema de organizar, actualizar y extraer información de una gran base de datos. Dreyfus denuncia que el problema del conocimiento no es un problema *per se,* sino sólo un problema artificial generado por las tesis del atomismo lógico (Dreyfus, 1992, p. 210). Y, dado que las computadoras sólo pueden operar sobre datos independientes y descontextualizados como los postulados por el atomismo lógico, la IA simbólica está condenada a afrontar esos problemas que ella misma produce.

El atomismo lógico, como teoría positivista que es, pretende descomponer la realidad en elementos independientes y descontextualizados, a imagen y semejanza de la física, con el objetivo de lograr un dominio mediante la prognosis tan eficaz como el que esta ciencia viene exhibiendo desde que en los tiempos de Galileo fue formalizada en ecuaciones matemáticas y adoptó el estilo explicativo nomológico-deductivo. La creencia de que todo puede ser tratado como conjuntos de elementos discretos se fundamenta, dice Dreyfus, en la confusión de los conceptos de mundo y universo (Ibíd., p. 213), un binomio que corresponde, respectivamente, a los conceptos de mundo social y mundo físico que venimos manejando desde el primer capítulo. Mientras que el *universo* (*universe*) es la totalidad de la materia, y ésta en principio sí es descriptible de manera atómica, el *mundo* (*world*), que es el entramado humano que proporciona significado a todo, incluido el universo, no es descriptible en términos atómicos como vamos a ver a continuación. De manera más precisa, Ramón Rodríguez

define así lo que es el mundo (*Welt*) desde el punto de vista de Heidegger, que es el compartido por Dreyfus: «La totalidad de relaciones de sentido, que forma parte de la propia existencia humana, y que actúa como el trasfondo a partir del cual una cosa determinada puede ser lo que es» (Rodríguez, 1987, p. 214).

A los distintos estados de cosas del universo los denomina Dreyfus *estados físicos* (*physical states*), mientras que a los estados de cosas del mundo se refiere como *situaciones* (*situations*), *contextos* (*contexts*) o *marcos* (*frames*). La correspondencia entre ambos no es unívoca, de tal forma que a un estado físico le correspondiese una y siempre la misma situación, sino que, muy al contrario, la misma situación puede darse respecto de estados físicos distintos, y un mismo estado físico puede ser interpretado desde situaciones diferentes en función de las metas y las intenciones de los seres humanos que participan en él (Dreyfus, 1992, p. 213).

Esto se aprecia claramente con un par de ejemplos. Supongamos que un maestro participa en dos partidas de ajedrez en las que llega a sendos momentos en que los tableros son idénticos: hay las mismas piezas y están colocadas de la misma manera. Este mismo estado físico puede, sin embargo, corresponder a dos situaciones bien distintas, pues el maestro no interpreta el tablero por sí solo, sino que su interpretación depende de otros factores tales como el estilo de juego que ha venido desarrollando su adversario a lo largo de la partida y de su carrera ajedrecística. Así, en la partida contra el oponente A, el maestro puede percibir una situación, mientras que contra el oponente B la situación puede ser otra. De hecho, una de las quejas de Kaspárov tras perder contra Deep Blue fue que no se le había permitido acceder al historial de partidas de la máquina, mientras que los ingenieros de IBM pasaron meses estudiando el historial de Kaspárov para deducir las claves de su estilo. Y, a la inversa, la misma situación puede corresponder a diferentes estados físicos. La situación para atacar enfilando las torres, por ejemplo, puede darse sobre una cantidad innumerable de combinaciones de piezas: puede hacer más o menos peones, estar situados una casilla más acá o más allá, y así con casi todas las piezas. Gracias a la perspicacia (*insight*) el maestro reconoce situaciones comunes ante estados físicos distintos.

Otro ejemplo, más natural que el del ajedrez, es el significado de la expresión "estar en casa". "Estar en casa" no puede definirse físicamente como estar dentro de los límites espaciales de un inmueble. Pensemos en la sentencia "Estuve en casa, *pero* estuve trabajando". La conjunción adversativa "pero" no tendría sentido si el estar en casa se definiera en términos puramente físicos. No obstante, el hecho es que sí tiene sentido, y lo tiene porque "estar en casa" refiere a un conjunto de prácticas, tales como dormir, comer y descansar, que son incompatibles con el trabajo. Por tanto, "estar en casa", antes que a un estado físico del universo, refiere a una situación dentro del mundo. Para enlazar este asunto con lo visto anteriormente, recordemos que en el capítulo quinto Schank señalaba que las estructuras memorísticas de todo tipo, principalmente los MOPs, suelen presentarse en tríos, conteniendo uno de ellos la *información física*, otro la *información social* y un último la *información personal* o *idiosincrásica* (Schank, 1999, p. 115). Schank ponía los siguientes ejemplos de MOPs contenedores de información física: «conducir, esperar, ser examinado, marcharse» (Ibíd., p. 124). Creer, como hace Schank, que "conducir", "esperar", "ser examinado" o "marcharse" son fenómenos descriptibles en términos puramente físicos es tan absurdo como creerlo respecto de "estar en casa". A todos ellos se les pueden poner un "pero", porque en realidad se trata de situaciones en el mundo, no de estados físicos. Por tanto, Schank incurre en la confusión de mundo y universo denunciada por Dreyfus al trazar una relación unívoca de correspondencia entre las situaciones de "conducir", "esperar", "ser examinado" y "marcharse" y unos estados físicos.

Hasta las expresiones que en apariencia refieren a estados físicos, como "estar en casa", "conducir", "esperar", "ser examinado" y "marcharse", refieren en realidad a situaciones en el mundo, y las situaciones, a diferencia de los estados físicos, no son descriptibles en elementos discretos, pues no tiene sentido preguntarse por cuáles son las partes independientes, atómicas, de las que se componen "estar en casa" o "conducir". Para mostrar lo absurdo que es llevar a cabo semejante descomposición Dreyfus escoge el ejemplo de la silla de Wittgenstein, y se pregunta por cuáles son las partes de las que se compone algo supuestamente tan físico como una silla. El segundo

Wittgenstein arremete de la siguiente manera en las *Investigaciones filosóficas* contra la extraña noción de "constituyentes más simples" del *Tractatus*: «¿Pero cuáles son las partes constituyentes simples de las que se compone la realidad? –¿Cuáles son las partes constituyentes simples de una silla? –¿Los trozos de madera con los que está ensamblada? ¿O las moléculas, o los átomos? –"Simple" quiere decir: no compuesto. Y aquí surge luego: ¿'compuesto' en qué sentido? No tiene ningún sentido hablar absolutamente de 'partes constituyentes simples de la silla'. [...] A la pregunta *filosófica*: "Es la figura visual de este árbol compuesta, y cuáles son sus partes constituyentes?", la respuesta correcta es: "Eso depende de qué entiendas por 'compuesto'". (Y ésta no es naturalmente una contestación sino un rechazo de la pregunta)» (Wittgenstein, 1953, §47).

El positivista y el ingenuo responderían que las partes de una silla son las patas, el asiento y el respaldo, pero en realidad no reconocemos un objeto como silla por ninguna de esas partes, pues siempre pueden faltar las patas y el respaldo, y la superficie horizontal que se supone que es el asiento puede ser en realidad una mesilla para depositar pequeños objetos decorativos. Interpretar que un objeto es una silla no depende por tanto de sus partes físicas, sino que significa entender su relación con otros objetos y con los seres humanos: «Esto implica todo un contexto de actividad humana del cual la forma de nuestro cuerpo, la institución social de los muebles, y la inevitabilidad de la fatiga constituyen sólo una pequeña parte. Y estos factores a su vez no son más aislables que la silla. Todos adquieren su significado en el contexto de la actividad humana de la que forman parte» (Dreyfus, 1992, p. 210). Así pues, la silla es una silla, como cualquier otra cosa se presenta como lo que se presenta, en virtud de su aparecer en una situación. «Sin embargo, las situaciones plantean problemas formidables para aquellos que desean traducirlas a un sistema formal» (Ibíd., p. 214). Explicitar y formalizar las prácticas sociales significadas por una silla o un cerdito-hucha, que es a lo que aspiran los proyectos de IA simbólica como el CYC, se presenta como una tarea extremadamente difícil. Para mostrar la dificultad de semejante formalización Dreyfus regresa al problema de la traducción automática.

Dada la frase "Él es un seguidor de Marx", ésta puede tener, al menos, dos significados: que es un seguidor de las ideas de Karl Marx, o que es un admirador de Groucho Marx. La manera de desambiguar es recurrir al contexto, pues no es lo mismo pronunciar esa sentencia en un seminario de filosofía que en un congreso de humoristas. El problema, por tanto, es cómo identificar el contexto. Así volvemos al problema de la selección del marco debido a la estructura circular de la comprensión: la identificación de un marco se realiza a partir de sus características esenciales, pero las características esenciales sólo destacan entre las demás sobre la base de un marco. Hay dos formas de habérselas con este círculo, dice Dreyfus (Ibíd., p. 55). Una es la de los seres humanos, la otra, la de las computadoras. Los seres humanos nos adentramos en el círculo mediante el *estar ya-siempre-en-una-situación* (*be always-already-in-a-situation*). «Es nuestro sentido de la situación lo que nos permite seleccionar de entre la infinidad potencial de hechos aquellos que son relevantes, y una vez esos hechos han sido encontrados, nos permite estimar su significado. Esto sugiere que, a menos que haya algunos hechos cuya relevancia y significado sean invariables en todas las situaciones –y nadie ha encontrado hasta ahora hechos así– tendremos que capacitar a las computadoras para reconocer situaciones; de lo contrario, no podrán desambiguar y por tanto serán, en principio, incapaces de entender proferencias en un lenguaje natural» (Ibíd., p. 218).

Mientras que los seres humanos entendemos la situación en términos del significado de las palabras tanto como entendemos el significado de las palabras en términos de la situación, las computadoras, como adelantábamos en el capítulo quinto, necesitan romper esta determinación recíproca, circular, y descomponerla en un proceso lineal de operaciones discretas (Ibíd., p. 220). Weizenbaum cree posible romper el círculo, dice Dreyfus, determinando primero el contexto, y a partir de él desambiguar el significado de las palabras. Sin embargo, la determinación del contexto requiere, al igual que la determinación del significado de cualquier palabra, de un elemento desambiguador que indique, primero, cuáles de entre los infinitos hechos que se presentan en todo momento son relevantes para identificarlo, y segundo, en

qué sentido han de ser interpretados esos hechos, pues los mismos hechos pueden ser relevantes para identificar varios contextos diferentes. Ese elemento desambiguador ha de ser, necesariamente, un contexto más amplio. Así, Weizenbaum asume el *supuesto metafísico* (*metaphysical assumption*) de que los contextos son manipulables como cualquier otro dato (Ibíd., p. 56) y por tanto pueden ser ordenados en una *estructura nido* como la de un árbol de decisión. Esto es algo que también está presente en la teoría cognitivista de la inteligencia de Schank, tal y como vimos en el capítulo quinto: interrelacionar jerárquicamente estructuras memorísticas. Los scriptlets, recordemos, apuntaban a escenas mediante etiquetas o índices, las escenas apuntaban a MOPs, y por último los MOPs apuntaban a TOPs, unas difusas estructuras memorísticas de aplicación intercontextual (Schank, 1999, p. 145).

Supuestamente, de ese modo reconoceríamos los contextos por sus rasgos esenciales, los cuales nos vendrían dados por el contexto inmediatamente superior en la jerarquía. Y así hasta llegar a un contexto último, definitivo (*ultimate*), que evitaría el regreso al infinito, y al que Weizenbaum denomina la *cultura compartida* (*shared culture*) (Dreyfus, 1992, p. 221). Ahora bien, dado que la computadora electrónica, como sistema formal que es, no es capaz de desambiguar el significado de ningún signo si no es por apelación a una regla o contexto de orden superior en la jerarquía, para detener el regreso al infinito es necesario que el contexto último posea rasgos que tengan un significado fijo, invariable, que le pertenezcan sólo a él (Ibíd., p. 289). Por tanto, así llegamos nuevamente a la necesidad de que el mundo tenga puntos de correspondencia unívoca con el universo, lo cual es falso como ya hemos visto, pues un mismo estado físico puede corresponder a varias situaciones del mundo y una misma situación puede darse en varios estados físicos, entendiendo los estados físicos en el sentido amplio del signo: ya sea el grafema o el fonema de una expresión como "estar en casa", un gesto con la mano o una forma geométrica.

La única solución pasa entonces por que el programador elija una situación última y la defina en términos de lo que suele ser relevante en ella: en la situación de jugar al ajedrez, es relevante la disposición de las piezas, en la de empaquetar tableros

de ajedrez, es relevante el tamaño y el peso. La consecuencia es que la máquina sólo es capaz de habérselas con esas situaciones predeterminadas *ad hoc*, y así padece la limitación de dominio que hemos observado en los sistemas expertos. Por el contrario, pretender que la máquina sea capaz de adaptarse a cualquier situación, como es propio de la inteligencia humana para hacer un intento pasable en casi cualquier cosa, conduce a una regresión de contextos que sólo se detiene en una descripción del mundo unívoca, como la del atomismo lógico, en la que se supone que hay signos simples que refieren a hechos elementales independientes y evidentes por sí mismos que no requieren interpretación.

7.1.3. Juegos sin reglas

En conclusión, al concebir la inteligencia como un proceso de cálculo, apunta Dreyfus, la IA simbólica desemboca en una *antinomia* (Ibíd., p. 222). Por un lado, la *tesis*: debe haber siempre un contexto más amplio para evitar la *circularidad* del problema de la selección del marco. Por otro, la *antítesis*: al deshacer la circularidad, debe haber un contexto último para evitar el *regreso al infinito*. La tercera posibilidad sería apelar al hecho fenomenológico antes mencionado de que el ser humano se encuentra ya-siempre-en-una-situación gracias a que la situación actual es una continuación o modificación de la anterior. Sin embargo, esta solución no es aplicable a las máquinas, pues eliminaría el *regreso al infinito jerárquico* a cambio de introducir un *regreso al infinito temporal*, el cual terminaría conduciendo a que debe haber una situación, contexto o marco primero comprensible por sí mismo a la manera del atomismo lógico. ¿Cuál es la solución entonces? Cambiar el planteamiento de raíz, dice Dreyfus: negar la separación entre hechos y contexto (Ibíd., p. 224). Negar que los hechos sean comprensibles con independencia de la situación. Simplemente, no tiene sentido hablar de hechos en sí mismos. No vemos un objeto que puramente está ahí y después lo interpretamos como una puerta, dice Heidegger, sino que directamente vemos una puerta (Heidegger, 1927, p. 173).

Una vez expuestos los argumentos epistemológicos de Dreyfus contra la posibilidad técnica de la IA simbólica, reflexionemos sobre este asunto basándonos en la filosofía pragmática del segundo Wittgenstein a la que él alude. Que los hechos dados a la conciencia se presenten siempre ya interpretados, como concluye Dreyfus, es asumible para una teoría psicológica como la fenomenología, pero la fenomenología no ofrece un modelo de la mente replicable por las computadoras, tal y como demuestran los cuatro procesos cognitivos examinados antes: periferia de la conciencia, tolerancia a la ambigüedad, discriminación de lo esencial y lo inesencial y agrupación perspicaz. Sólo hay dos modelos de la mente realizables por la IA simbólica. Uno, el de la psicología cognitiva, debido a que se basa en el supuesto nuclear, y por tanto irrenunciable, de que la mente es en realidad un programa informático, que es lo afirmado por el supuesto psicológico y la HFSS. Y dos, el de cualquier psicología que presente un modelo de la mente que, aunque no sea un programa informático, su funcionamiento sea descriptible como un programa informático, que es lo afirmado por el supuesto epistemológico y la HSS. Por tanto, lo mínimo indispensable que ha de tener un modelo de la mente para ser realizable por la IA simbólica es que sea descriptible como un programa informático.

Los programas informáticos, como vimos en el capítulo tercero, son sistemas formales, y los sistemas formales se componen de un lenguaje formal y un mecanismo deductivo (Falguera & Martínez, 1999, p. 64). A su vez, un lenguaje formal es un lenguaje artificial exento de interpretación semántica en su definición que se compone de un vocabulario y una sintaxis, siendo la sintaxis un conjunto de reglas para formar expresiones o fórmulas bien formadas con el vocabulario. En cuanto al mecanismo deductivo, se compone, como mínimo, de un conjunto de reglas de transformación que sirven para transformar unas expresiones en otras, y en ocasiones también de un conjunto de axiomas, que son proposiciones de partida evidentes. Un sistema formal es, en resumen, un conjunto de *elementos* y otro de *reglas* que determinan las relaciones entre ellos: éstos son, por tanto, los términos en los que ha de ser descriptible un modelo de la mente para ser realizable por la IA simbólica.

Por otra parte tenemos que el positivismo, guiado por el deseo de dominación mediante la razón instrumental, considera que las únicas explicaciones científicas, y como tales las únicas plenamente satisfactorias, son aquellas que reducen los fenómenos a elementos relacionados por reglas. La consecuencia es que, desde el punto de vista del positivismo, si la psicología puede ser una ciencia, entonces la posibilidad teórica de la IA simbólica está garantizada. En este punto el positivismo se encuentra en una encrucijada. Ha de decidir si la psicología puede o no puede ser una ciencia. Para tomar la decisión tiene argumentos a favor y en contra. A favor está la ley de los tres estados del desarrollo humano enunciada por el padre del positivismo, Augusto Comte, según la cual la humanidad se encuentra actualmente en el estado positivo, caracterizado por un avance continuo del conocimiento científico que no se detendrá hasta reemplazar a todas las viejas creencias teológicas y metafísicas (Comte, *Curso de filosofía positiva*, p. 21). En este sentido, es deseable que la psicología sea una ciencia, para quer silencie las habladurías de los sacerdotes y de los filósofos acerca del alma. Pero, por el otro lado, en contra de que la psicología pueda ser una ciencia, está la advertencia materialista también de Comte de que el único conocimiento real es aquel que se basa en los hechos observados, y los fenómenos mentales, aunque son observables, sólo lo son a través de la introspección, la cual es un método de observación defectuoso debido a que el sujeto y el objeto observado son la misma persona, y eso produce una deformación de la experiencia (Ibíd., p. 49).

El resultado de esta encrucijada de la psicología para el modo de pensar positivista ya lo conocemos. A principios del siglo XX fue el conductismo, y a mediados, el cognitivismo junto con su consecuencia necesaria derivada de su noción de lo que es explicar: la IA simbólica. El paradigma cognitivista surgió como reacción contra el conductismo, pero compartiendo con éste la misma pretensión positivista de explicarlo todo en términos de elementos atómicos y reglas para convertir a la psicología en una ciencia instrumentalmente tan eficaz como cualquier ciencia natural (Haugeland, 1978, p. 279). Los primeros cognitivistas como George Miller y Ulric Neisser reprochaban al conductismo que sus leyes de la conducta, aunque eficaces, no

servían para explicar las conductas organizadas complejas, siendo el lenguaje la más importante. La ley del condicionamiento operante, formulada por Skinner, permite una predicción y una manipulación de los sujetos muy cercana a la predicción y la manipulación que las leyes de Newton permiten sobre la materia. El propósito del cognitivismo es descubrir leyes tan eficaces como las del condicionamiento operante y el condicionamiento clásico, pero sobre conductas más complejas.

El problema de las conductas más complejas es que su nivel de complejidad es directamente proporcional al número de factores que intervienen. Herbert Simon descubrió esta relación al reflexionar sobre el camino trazado por una hormiga sobre las dunas de una playa: «Vista como una figura geométrica, la trayectoria de la hormiga es irregular, compleja, difícil de describir. Pero su complejidad es en realidad la complejidad de la superficie de la playa, no una complejidad interna de la hormiga. En esa misma playa, otra criatura pequeña con una meta en el mismo lugar que la hormiga seguiría un camino muy similar» (Simon, 1981, p. 51). En el caso de la inteligencia humana sucede lo mismo, pues observamos que las conductas más complejas como descubrir teorías científicas o realizar obras de arte suelen ser producidas por sujetos que en su desarrollo han estado inmersos en entornos de estímulos ricos, complejos. Y, a la inversa, es infrecuente, por desgracia, que el hijo de un chatarrero, que se ha criado entre basura, llegue a ser un gran intelectual.

Por tanto, cuanto más compleja es una conducta, más complejos son los factores de los que surge. En una conducta simple como la de respuesta asociativa por condicionamiento operante los factores son pocos. Tan pocos, que se pueden controlar experimentalmente en la caja diseñada por Skinner: un lugar estanco, aislado del resto del universo, en el que sólo hay comida, agua, un timbre, una luz, una palanca, una rejilla electrificada. Ésos son todos los factores, elementos, estímulos o variables, según cómo se los quiera llamar. En cambio, el lenguaje es una conducta que jamás puede obtenerse introduciendo a un sujeto en un lugar así, porque el lenguaje es resultado de la inmersión desde la lactancia en un entorno repleto de una enorme variedad de estímulos físicos y sociales.

Subrayemos que es para esto para lo que surge el cognitivismo: para explicar conductas complejas, deliberativas como el lenguaje y no puramente reactivas. Sin embargo, sus resultados no son los prometidos. Lo único que el cognitivismo es capaz de explicar son fenómenos igualmente contenidos en recintos como la caja de Skinner. El ajedrez es el mejor ejemplo de tarea intelectual compleja explicable por el cognitivismo mediante sus programas informáticos de IA simbólica. Es un juego complejo, pero reparemos en que toda su complejidad consiste en las muchas posibilidades combinatorias de sus elementos. En cuanto a la cantidad de reglas y elementos no hay complejidad alguna: 64 casillas y 32 piezas de sólo 6 tipos que se relacionan entre sí por unas reglas tan simples que caben en media página. ¿Qué es el ajedrez si no una abstracción, una pequeña caja de Skinner?

Otros ejemplos de tareas explicables por el cognitivismo los encontramos en los sistemas expertos: analizar muestras de sangre, configurar ordenadores. En todos ellos los programadores determinan de antemano cuáles son los estímulos relevantes, o lo que es igual, los datos que deben entrar en la caja, y las reglas que hay que aplicarles para transformarlos en los datos de salida propios de una conducta inteligente. Es verdad que entran más datos que en el ajedrez y que en la caja de Skinner, por lo que podríamos decir metafóricamente que la caja es más grande, pero sigue siendo una caja, un micromundo. En cuanto se le pide al cognitivismo, o a su contrapartida experimental la IA simbólica, que salga de esas cajas al mundo real, entonces empiezan los problemas, porque en el mundo real hay demasiados estímulos, demasiados datos como para formalizarlos e introducirlos uno por uno en un dispositivo que, como vimos en el capítulo tercero, es por definición autocontenido en tanto que digital. Y lo que es aún más grave para el concepto positivista de explicación: muchos de ellos se relacionan entre sí en la mente sin que sus relaciones estén regidas por reglas ni sean descriptibles por reglas.

A través del lenguaje los seres humanos descomponemos el mundo en elementos, unos más simples y otros más complejos, y los designamos mediante expresiones como "estar en casa", "amistad" o "imperialismo". Los significados de las

palabras, dice Wittgenstein al comienzo de las *Investigaciones filosóficas*, no suelen ser referentes externos al lenguaje en el sentido primitivo de cosas señalables con el dedo como pensaba Agustín de Hipona. La mayoría de las palabras debe su significado a su participación en juegos de lenguaje. A diferencia de los juegos artificiales como el ajedrez, los juegos de lenguaje no son explicables en sentido positivista porque, detrás de unas cuantas capas de reglas, se descubre que su validez depende de condiciones dictadas por prácticas sociales totalmente arbitrarias, que no obedecen a reglas.

Recordemos un fragmento de la definición de Charniak de "cerdito-hucha": «Los cerditos-hucha (CH en adelante) los hay de todos los tamaños y formas, aunque la forma preferida es la de cerdito. Generalmente (*generally*) el tamaño oscila entre el del pomo de una puerta y una tartera. Generalmente (*generally*) en los CH se guarda dinero, por lo que cuando un niño necesita dinero a menudo irá a mirar en su CH. Usualmente (*usually*) para conseguir el dinero necesitas sostenerlo y agitarlo (de arriba a abajo). Generalmente (*generally*) ponerlo boca abajo facilita las cosas» (Crevier, 1993, p. 113). Observamos que el significado de "cerdito-hucha" es, aparte de una breve descripción física, un extenso conjunto de prácticas sociales codificadas en reglas. Ahora bien, estas primeras capas de reglas van precedidas de adverbios como "generalmente" o "usualmente" que advierten de que su cumplimiento no es universal, sino que hay excepciones.

La presencia de excepciones en las reglas que definen lo que es un "cerdito-hucha" no es un caso aislado, sino un reflejo de la norma general. Isaac Asimov, que además de pensar mucho sobre robótica era doctor en filosofía, abordó el tema de la excepcionalidad universal de las reglas humanas en el siguiente fragmento de su novela *Robots e Imperio*, en el que dos robots conversan, y uno le dice al otro: «Estoy seguro que en el largo recuento de los acontecimientos humanos debe de haber, escondidas, unas leyes para la humanidad equivalentes a las Tres Leyes de la robótica. [...] Pero aunque tengo la impresión de que estas leyes de la humanidad deben existir, no puedo encontrarlas. Cada generalización que intento plantear, por más amplia y sencilla que sea, contiene numerosas excepciones» (Asimov, 1985, p. 66).

La computadora puede operar con excepciones a las reglas, pero sólo si se le indica cuáles son esas excepciones y cómo identificarlas, pues de lo contrario aplica las reglas indiscriminadamente. Es lo apuntado por John McCarthy y Patrick Hayes con el problema de la cualificación. La dificultad radica en que los seres humanos no sabemos explícitamente cuáles son las condiciones de validez *ceteris paribus* de las reglas, o lo que es igual, cuándo se producen las excepciones. Retomando el ejemplo del regalo de cumpleaños, las condiciones de validez de la regla "*generalmente* nadie quiere tener dos objetos iguales" son innumerables. Pero la computadora las demanda, comportándose de manera tan exasperante como el niño que pregunta sin cesar el porqué del porqué. Suponiendo que el adulto que la programa tenga paciencia y siga el juego, al final se queda siempre sin respuestas: simplemente es así. Llegados a ese punto, al cognitivismo y a la IA simbólica sólo les queda aferrarse a la hipótesis de que, en el fondo, tras ese "simplemente es así" se esconde un proceso mental inconsciente que también es formalizable en reglas. *Debe* serlo. Dado que explicar es formalizar en reglas, para que la mente sea explicable, tiene que ser formalizable. Hay dos formas de acometer dicha formalización ineludible para la forma de pensar del positivismo: introduciendo las reglas con sus condiciones de validez una por una, o enseñando a la máquina a abducirlas por sí misma.

La primera vía es la del CYC, un fracaso importante en la Historia de la IA. Los ingenieros del CYC tienen el cometido de identificar todos los elementos de los que se compone el mundo y definirlos de manera semejante al cerdito-hucha, explicitando las reglas de las prácticas sociales en las que participan y sus condiciones de validez. La tarea es de tal magnitud que la estrategia de la psicología cognitiva es, como vimos en el capítulo cuarto, pedir que se le permita empezar por formalizar aquellos procesos mentales independientes de los factores más complejos: «los afectos, el contexto, la cultura y la historia» (Gardner, 1985, p. 58). A todos estos hay que sumarle la petición de excluir también la biología del cerebro y del cuerpo, pero no al comienzo de su andadura como nuevo paradigma para después incorporarlos progresivamente, sino por principio, por puro dogmatismo sin ningún indicio empírico que avale la analogía

de la mente con un programa informático y la supuesta división cartesiana, falsa por otra parte (Ceruzzi, 1998, p. 80), entre hardware y software que se da en las computadoras electrónicas. En la actualidad, transcurrido ya más de medio siglo desde su fundación, la psicología cognitiva sigue sin explicar los procesos mentales superiores como el lenguaje. El silogismo es simple: si los hubieran explicado, entonces habría inteligencias artificiales fuertes capaces de superar el test de Turing, pero no las hay porque las categorías especiales de los premios Loebner siguen desiertas, luego no los han explicado. No obstante, hay que reconocerle a Douglas Lenat, el director original del CYC, el valor de haberse lanzado a intentar formalizarlo casi todo. Él ha fracasado, pero al menos ha hecho frente a los factores ambientales más complejos.

La otra manera de formalizar la mente es la de descubrir las reglas por las cuales ésta produce sus propias reglas o teorías acerca de cómo funciona el mundo, tanto físico como social, dando siempre por supuesto que la mente opera al menos de una forma descriptible por reglas. Dreyfus no se ocupa de examinar esta vía, sino que, como él dice, sus argumentos están dirigidos sólo a las computadoras inteligentes que, como el CYC, han sido creadas al estilo de Atenea (Dreyfus, 1992, p. 290), que nació de la frente de Zeus con su forma adulta y pertrechada con sus armas de diosa de la guerra. En cambio, nosotros, al hablar del mito del método científico en el capítulo cuarto, ya expusimos argumentos contra la formalización algorítmica del proceso de producción de teorías, tanto científicas como precientíficas. Las estrategias de descubrimiento de teorías son, recordemos, lógicas y psicológicas.

Las lógicas son la inducción, la abducción y la preducción, y las psicológicas son la serendipia y la analogía. La serendipia, evidentemente, no puede ser programada en una computadora, pues por definición depende del azar. Respecto a la analogía, es programable en una IA simbólica, al igual que la inducción y la preducción, pero en última instancia estas tres estrategias desembocan en el problema de la selección del marco, pues requieren de una teoría previa. En el caso de la preducción, la razón es obvia, porque preducir no es más que derivar unas teorías de otras. Y en el de la inducción y la analogía, aunque no es tan obvio, no es menos verdadero, en tanto que

toda inducción o analogía se realiza sobre la base de una carga teórica que dirige la mirada a unos hechos y no a otros. Por ejemplo, la famosa inducción "Todos los cisnes son blancos" sólo es posible gracias a una teoría que de las observaciones de los cisnes abstrae a éstos y a su color como hechos relevantes, al tiempo que descarta todos los demás, tales como el paisaje, el tamaño del animal o la hora del día en la que se realiza la observación. Hay inteligencias artificiales simbólicas que, como decimos, son capaces de elaborar teorías mediante inducción y analogía, pero sus teorías, a causa del problema de la selección del marco, siempre estarán limitadas a dar cuenta de los hechos seleccionados como relevantes de antemano por el programador, y por tanto las máquinas de este tipo nunca serán lo suficientemente generales, sino que padecerán el mal de la limitación de dominio. Así como en la elaboración de teorías científicas la relevancia es determinada por el paradigma, a nivel precientífico es determinada por el marco o contexto. En el capítulo quinto ya mencionamos que, según Minsky, la función de los marcos en la actividad precientífica es análoga a la de los paradigmas en la científica: «Ahora bien, mientras que Kuhn prefiere aplicar su muy efectivo paradigma de redescripción al nivel de las grandes revoluciones científicas, a mí me parece que la misma idea se adapta bien al microcosmos del pensamiento cotidiano» (Minsky, 1975, p. 121).

En cuanto a la abducción, es la principal estrategia lógica de producción de teorías: dado un fenómeno se abstraen los factores que intervienen y se proponen las reglas que los relacionan. Por ejemplo, dado el fenómeno de la fuerza, Newton abdujo que ésta era directamente proporcional a la masa y la aceleración, formulando así su segunda ley: $F=m\cdot a$. De igual manera, a nivel precientífico también abducimos teorías para explicar la conexión entre distintos hechos. Sin embargo, en el capítulo quinto vimos que Schank, desde el paradigma cognitivista de la psicología, era incapaz de definir el algoritmo de la abducción. La causa técnica que impide formalizar algorítmicamente la abducción es la misma que la de la inducción, la preducción y la analogía: la necesidad de una teoría previa que, en el caso de la segunda ley de Newton, fije la atención en la masa y la aceleración, y no en el color o la forma. Sin

seleccionar previamente las características relevantes, todos los objetos comparten un listado indefinidamente largo de relaciones, pero dicha selección depende de la selección previa de una teoría, que es justo lo que se busca, por lo que al final se cae en un círculo sólo solucionable mediante un regreso al infinito no menos absurdo.

A Einstein no le habría sorprendido el fracaso de Schank y de la IA simbólica en su intento por formalizar la abducción, pues ya mencionamos en el capítulo cuarto un texto en el que el físico alemán declaraba su convencimiento de que no existe ningún camino lógico, ningún algoritmo, para abducir teorías: «No hay camino lógico que lleve a estas leyes fundamentales. Debemos dejarnos conducir por la intuición, que se basa en una sensación de la experiencia. [...] Nadie que haya profundizado de veras en esto podrá negar que el sistema teórico ha sido prácticamente determinado por el mundo de las suposiciones, pese a que no existe camino lógico alguno que conduzca desde éstas hasta las leyes fundamentales» (Einstein, 1955, p. 131).

El mito del método científico y el mito de la IA simbólica se basan en la misma creencia positivista de que la mente es explicable por reglas, pues ambos sostienen que existe un algoritmo para producir teorías, científicas en un caso y precientíficas en el otro. En consecuencia, se pueden utilizar los argumentos contra la posibilidad del método científico para argumentar contra la posibilidad de la IA simbólica, y a la inversa, dado que la IA simbólica es imposible, el método científico también lo es. El método científico no es ni podrá ser jamás un algoritmo de producción de teorías, sino tan sólo lo que ya es en la actualidad: un criterio para distinguir entre teorías científicas y precientíficas. El criterio consiste en *protocolos* que establecen cómo ha de dirigirse la actividad para evitar los efectos distorsionadores de ciertos factores que, si bien participan en la producción de teorías precientíficas, en la producción de las científicas son inadmisibles. O, dicho de otro modo, los protocolos son cajas de Skinner virtuales que pretenden encerrar las mentes de los científicos en micromundos de estímulos controlados, pero no dicen cómo hacer lo que hay que hacer con lo que sí entra ahí: producir nuevas teorías, ya sea por inducción, abducción, preducción, serendipia o analogía. No hay, por tanto, ningún método para producir conocimiento.

La creencia de que existe un método que garantiza el avance imparable de la ciencia es sostenida por dos agentes con sus respectivos intereses. Por un lado, los propios científicos, para que la sociedad invierta en sus proyectos sin temor a estar malgastando el dinero en horas de laboratorio sin resultado. Véase si no la fiereza con la que reclaman inversiones públicas en ciencia, como si ese dinero no fuera a producir artículos tan fraudulentos como los denunciados en 2012 por la PNAS (*Proceedings of the National Academy of Sciences*) en el estudio al que ya nos referimos en el primer capítulo (Fang, Steen & Casadevall, 2012). Recordemos su conclusión: en el campo de la biomedicina y las ciencias de la vida el número de artículos fraudulentos se ha multiplicado por diez desde 1975. De un seguimiento de 2.047 artículos retractados se desprendía que sólo un 21,3% fueron atribuibles a errores, mientras que el 67,4% restante se debía a mala conducta. Dentro de los artículos producidos por mala conducta, el 43,4% había sido retirado por fraude o sospecha de fraude, el 14,2% eran duplicaciones, y el 9,8% eran directamente plagios. Y todo ello habiendo un supuesto método científico para dirigir bien la razón y buscar la verdad en las ciencias, lo cual es tranquilizador, pues da la impresión de que, si no lo hubiera, algunos investigadores mentirían hasta en sus juicios realizativos puros como el dar los buenos días.

Pero de nada sirven las revelaciones de este tipo, porque la sociedad, que es el otro agente que sostiene la creencia en el método científico, quiere creer que existe un tal método que garantiza la transformación del dinero en descubrimientos con sus consiguientes aplicaciones técnicas para mejorar la calidad de vida. Ortega y Gasset denuncia así esta ilusión: «¿Se ha pensado en todas las cosas que necesitan seguir vigentes en las almas para que pueda seguir habiendo de verdad "hombres de ciencia"? ¿Se cree en serio que mientras haya *dollars* habrá ciencia? Esta idea en que muchos se tranquilizan no es sino una prueba más de primitivismo» (Ortega y Gasset, 1930, p. 103). Ortega habla de algo tan inaprensible como las "almas" de los hombres de ciencia y de todas las cosas que tienen que seguir vigentes en ellas para que produzcan teorías. Lo que tiene que seguir vigente es todo un mundo de estímulos, y no uno cualquiera, pues como señala el genial filósofo español: «En toda la amplitud

de la tierra y en toda la del tiempo, la fisicoquímica sólo ha logrado constituirse, establecerse plenamente en el breve cuadrilátero que inscriben Londres, Berlín, Viena y París. Y aún dentro de ese cuadrilátero, sólo en el siglo XIX. Esto demuestra que la ciencia experimental es uno de los productos más improbables de la Historia. [...] Esta fauna del hombre experimental requiere por lo visto, para producirse, un conjunto de condiciones más insólito que el que engendra un unicornio. Hecho tan sobrio y magro debía hacer reflexionar un poco sobre el carácter supervolátil, evaporante, de la inspiración científica. ¡Lucido va quien crea que si Europa desapareciese podrían los norteamericanos *continuar* la ciencia!» (Ibíd., p. 130).

El primitivismo que sustenta la confianza en el avance continuo de la ciencia es el mismo que sustenta la confianza en la IA simbólica a pesar de sus evidentes síntomas de programa de investigación degenerado. Se desea que la mente sea explicable por reglas como cualquier parcela de la naturaleza para dominarla de la misma manera que se domina la tierra mediante la agricultura, los microorganismos mediante la biología y la energía nuclear mediante la física. Contra la ansiedad, haga esto. Contra la depresión, esto otro. Si quiere ser feliz, en definitiva, manipule su mente como indica el algoritmo. Para curarse de un mal pensamiento sólo habría que querer hacerlo, y la ciencia de la psicología aplicaría sus leyes para lograrlo. La mente sería tan manipulable como cualquier sustancia material. Así la voluntad conquistaría su último bastión, pues no habría nada que no pudiera conseguirse a fuerza de ella, a fuerza de voluntad. Sobre este tipo de cuestiones, acerca de las condiciones de posibilidad sociales de la IA, seguiremos hablando en el próximo capítulo.

Con las reflexiones sintetizadoras de estas últimas páginas, que abarcan varios asuntos analizados en los capítulos previos, damos por finalizado el examen de las condiciones de posibilidad técnicas de la IA simbólica en sentido fuerte. La Historia de la IA simbólica es la historia de unos hombres de ciencia que, a pesar de su indudable ingenio, se embarcaron en un proyecto imposible por no haber filosofado antes. En los años 50, mientras que Wittgenstein ya se había retractado de su *Tractatus*, los investigadores de la IA simbólica, deseosos, como es natural, de creer en la viabilidad

de su proyecto de construir computadoras electrónicas inteligentes, se vieron obligados, y todavía siguen estándolo, a adoptar una filosofía positivista como la del *Tractatus*, en la que mundo, lenguaje y pensamiento son isomorfos, siendo su forma verdadera común la de un sistema formal, es decir, la de un programa informático. Veamos a continuación, también desde el punto de vista filosófico de la epistemología, las condiciones de posibilidad técnicas de la IA subsimbólica.

7.2. Problemas de la IA subsimbólica

A diferencia de la IA simbólica, la IA subsimbólica no necesita postular una hipótesis tan cuestionable como la de la existencia de representaciones mentales cuasi-lingüísticas manipulables de acuerdo a reglas sensibles a la estructura. Su fundamento, en cambio, es un hecho bien contrastado. Dado que su objetivo es replicar la conducta inteligente mediante la duplicación realista o instrumental del funcionamiento de las redes de neuronas del cual surgen tanto la mente como la conducta, la IA simbólica se basa en el hecho indiscutible de que mente y conducta surgen del cerebro (Hawkins & Blakeslee, 2004, p. 59). En consecuencia, la única condición de posibilidad técnica que debe cumplir la IA subsimbólica es replicar el cerebro con *suficiente* detalle (Franklin, 1995, p. 122). Esto es lo que en el capítulo anterior denominamos la *hipótesis de la semejanza aproximada* (*rough resemblance hypothesis*), o HSA.

Cuál debe ser ese grado de semejanza es una cuestión empírica. Bien pudiera ser que fuese necesario replicar hasta el funcionamiento individual de los canales iónicos, o quizás no. Es una disyuntiva que sólo puede dirimirse experimentalmente, y que por tanto queda fuera de nuestro alcance. Lo que sí podemos hacer aquí, y va a ser el tema de esta sección, es señalar las diferencias que en la actualidad separan a las redes de neuronas artificiales de las naturales. Nos centraremos exclusivamente en las redes de neuronas de la IA ajena, debido a que son las únicas con utilidades de IA fuerte, que es el enfoque de la IA que nos interesa. A día de hoy, las de la corriente de

la IA humana, como Blue Brain y las simulaciones antes mencionadas del equipo de Ananthanarayanan, sólo tienen el propósito científico de IA débil de servir para estudiar ciertos procesos neuronales (Santos & Duro, 2005, p. 92).

Diferencias principales

La cantidad de diferencias entre las redes de neuronas artificiales y las naturales es, en principio, infinita. Hay tantas diferencias entre ambos tipos de redes como diferencias hay entre el modelo de un objeto y el objeto mismo, pues las redes de neuronas artificiales son modelos de las naturales, y es consustancial a todo modelo que sea una simplificación de la realidad figurada por él (García, 2001a, p. 78). Las diferencias que aquí vamos a analizar son las señaladas por David Rumelhart y Paul Churchland. Se trata, por tanto, no de diferencias cualesquiera, sino escogidas por dos expertos en pensar sobre la IA subsimbólica atendiendo a su posible relevancia para lograr una duplicación instrumental eficaz del cerebro. Estas diferencias son: dependencia de la capacidad para generalizar respecto del número de unidades ocultas, número de conexiones de entrada y salida, ausencia de conexiones horizontales, variabilidad de efecto de las sinapsis, variedad de arquitecturas y tres diferencias relativas al aprendizaje que son la regla delta generalizada, la velocidad de adquisición y la velocidad de ejecución.

En el capítulo anterior dedicamos varias páginas a describir la estructura, el funcionamiento y las propiedades de los sistemas conexionistas estándar de tres capas. Éstos, recordemos brevemente, están formados por tres capas de unidades conectadas unidireccionalmente en este orden: unidades de entrada, unidades ocultas y unidades de salida. Una de las propiedades más importantes de estos sistemas, dijimos, es la capacidad de producir generalizaciones basadas en la similaridad, con las cuales responder exitosamente a situaciones inéditas pero similares a las anteriormente experimentadas. El filósofo Paul Churchland señala que la capacidad para *generalizar* depende del *número de unidades ocultas* (Churchland, 1989, p. 278).

Si no son suficientes, es imposible que aprendan a generalizar, y si son demasiadas, el entrenamiento requerido para enseñarlas se prolonga en exceso. La razón es que, si no son suficientes, no serán capaces de almacenar una representación interna del prototipo subyacente a las entradas proporcionadas. Si, por el contrario, son demasiadas, lo único que hará el sistema es memorizar en ellas representaciones particulares de cada entrada sin relacionarlas entre sí. Para que se produzca la generalización es necesario alcanzar una masa crítica de experiencias cuyo número es directamente proporcional al de unidades ocultas. En el cerebro, en cambio, no sucede nada parecido. Entre las neuronas sensoriales y las motoras hay una gran cantidad de interneuronas cuya cantidad no necesita ser ajustada externamente. De hecho, la inmensa mayoría de las neuronas del sistema nervioso son interneuronas. Por tanto, he aquí la primera diferencia importante, y es que las redes de neuronas artificiales requieren de un ajuste estructural *ad hoc* realizado por un agente externo al sistema.

En cuanto al *número de conexiones de entrada y de salida*, son dos conceptos a los que en el capítulo anterior dimos los nombres respectivos de fan-in y fan-out (Rumelhart, 1989, p. 213). En los sistemas conexionistas estándar cada una de las unidades de entrada proyecta sobre todas las unidades ocultas, y cada una de éstas proyecta a su vez sobre todas las unidades de salida. Esto es algo que no sucede en el cerebro humano, pues teniendo 10^{11} neuronas tan sólo cuenta con 10^{14} sinapsis, es decir, una media de 1.000 sinapsis por neurona (Kandel, Schwartz & Jessell, 1995, p. 183). Si extrapolásemos la estructura de los sistemas conexionistas al sistema nervioso, resultaría que éste debería tener tantas conexiones por neurona como neuronas por capa. Dado que la corteza consta de $3x10^{10}$ neuronas divididas en seis capas, cada una de esas neuronas debería proyectar $5x10^{9}$ sinapsis, una cantidad astronómica incompatible con el fan-in máximo del cerebro, que se cifra en las 150.000 sinapsis recibidas por cada célula de Purkinje en el cerebelo (Kandel, Schwartz & Jessell, 1995, p. 27). Respecto a las *conexiones horizontales* entre neuronas de una misma capa sucede lo contrario, y es que mientras que en el cerebro son abundantes, en los sistemas conexionistas estándar no las hay (Churchland, 1989, p. 281).

En el capítulo quinto vimos que, atendiendo a su morfología, las sinapsis entre neuronas pueden ser axodendríticas, axosomáticas y axoaxónicas, es decir, de axón a dendrita, de axón a soma y de axón a axón. Por lo general, en el sistema nervioso central las axodendríticas son excitatorias, las axosomáticas son inhibitorias y en las axoaxónicas el axón presináptico modula la sinapsis del axón postsináptico. En cualquier caso, lo que nos interesa aquí es que las sinapsis *excitatorias* no pueden transformarse en *inhibitorias*, y viceversa, las inhibitorias no pueden pasar a ser excitatorias. Es algo que no se puede cambiar mediante ninguna alteración química, pues la acción excitatoria o inhibitoria de las sinapsis químicas, que son las más abundantes y las únicas capaces de producir inhibición, no depende de los neurotransmisores liberados, sino de las propiedades de los receptores de la neurona postsináptica que los captan (Kandel, Schwartz & Jessell, 1995, p. 192). En cambio, en las redes de neuronas artificiales es necesario que una conexión sea capaz tanto de inhibir como de excitar (Churchland, 1989, p. 281). Así, una conexión excitatoria, por ejemplo, tiene que poder convertirse en inhibitoria por efecto del entrenamiento. Sin esta plasticidad irreal, los sistemas conexionistas estándar no son funcionales (Copeland, 1993, p. 334). Esta diferencia entre las redes de neuronas artificiales y las naturales, señala Churchland, se puede remediar alterando la arquitectura básica de tres capas, simplemente añadiendo interneuronas inhibitorias. Esta observación nos lleva al siguiente punto, el de la variedad de arquitecturas.

Las diferencias apuntadas hasta ahora son sólo aplicables, como venimos advirtiendo, a los sistemas conexionistas estándar de tres capas. Sin embargo, existen otras arquitecturas alternativas. A modo de ejemplos, Rumelhart menciona las arquitecturas de Mike Jordan y Geoffrey Hinton (Rumelhart, 1989, p. 229). La de Mike Jordan consiste en cuatro capas de unidades: unidades de planificación, de contexto, ocultas y de salida. Las de planificación proyectan sobre las ocultas, que a su vez proyectan sobre las de salida, que finalmente proyectan sobre las de contexto para retroalimentar las ocultas. La utilidad de esta arquitectura es la de producir secuencias de fonemas. En cuanto a la arquitectura de Hinton, consiste en una variante de la de

tres capas, con la peculiaridad de que las capas de entrada y de salida son idénticas. Estos son sólo algunos ejemplos de la infinidad de configuraciones arquitectónicas que puede adoptar una red de neuronas artificiales. Por su parte, las naturales también tienen diferentes arquitecturas, pero están dotadas de una gran flexibilidad que permite su reorganización automática para realizar casi cualquier tarea. Gracias a esa flexibilidad es posible, como mencionamos en el capítulo quinto, que la corteza visual primaria de un sujeto que se ha quedado ciego pase a desempeñar labores de procesamiento de información táctil. Por tanto, mientras que las redes de neuronas artificiales tienen arquitecturas particulares, las redes de neuronas naturales están dispuestas de tal forma que podrían ser calificadas de universales.

Respecto al aprendizaje, una manera muy eficaz, aunque no la única, de entrenar sistemas conexionistas es la *regla delta generalizada* descubierta por Rumelhart, Hinton y el resto del Equipo PDP. Esta regla, como ya dijimos en el capítulo anterior, consta de dos fases (Ibíd., p. 226). En la primera se le aplica una entrada a la red, y en la segunda se compara la salida producida con el resultado deseado. La diferencia es denominada *error*. Si el error es nulo o aceptablemente bajo, entonces la red está bien ajustada. De lo contrario, el error es enviado hacia atrás en la jerarquía, en dirección a la capa de entrada, para modificar los pesos de las conexiones de entrada de las unidades de la segunda capa, y así sucesivamente hacia capas anteriores. Una vez se han hecho los ajustes pertinentes en base al error, se introduce una nueva entrada en la red y se repite el ciclo. El problema de este proceso, denominado propagación hacia atrás del error, es que requiere que un observador externo determine cuál es el resultado deseado (Churchland, 1989, p. 282). Esto es algo innecesario en los seres vivos, que son capaces de aprender la mayor parte del tiempo sin que nadie les diga si su conducta es correcta.

Otra razón adicional por la que la regla delta generalizada, a pesar de su eficacia, no apunta hacia el mecanismo real de aprendizaje utilizado por el cerebro es que las redes de neuronas naturales aumentan la *velocidad de ejecución* a medida que aumenta el entrenamiento. Sin embargo, con la regla delta generalizada este efecto no

se observa. Aumenta la tasa de acierto, pero no la velocidad. En cuanto a la *velocidad de adquisición* de representaciones internas para mejorar la tasa de acierto, es la última de las diferencias que hemos enumerado. Hay otras reglas de aprendizaje alternativas a la regla delta generalizada que no requieren ninguna medida del error, y que por tanto no dependen de que un sujeto externo decida si la respuesta producida se desvía o no de la deseada (Santos & Duro, 2005, p. 93), pero ninguna de ellas soluciona el problema de la lentitud del aprendizaje de los sistemas conexionistas.

Los seres humanos aprendemos muy rápido. Y cuanto más rápido, más inteligentes se dice que somos. Habitualmente, con tan sólo unos cuantos ejemplos de muestra, somos capaces de inferir la regla general que los conecta. Es uno de los pilares de los tests de inteligencia: evaluar la rapidez con la que un sujeto abstrae normas correctas. En cambio, los sistemas conexionistas necesitan de un entrenamiento muy extenso, con una gran cantidad de ejemplos, y que además deben ser lo suficientemente desemejantes entre sí, pues de lo contrario las generalizaciones resultantes son defectuosas en tanto que demasiado estrechas, poco generales. Como dice Churchland: «Una red es siempre esclava de la calidad del muestreo de entrenamiento en relación a la población total» (Churchland, 1989, p. 280). Así, por ejemplo, si en un muestreo para entrenar a una red en el reconocimiento de cisnes todos los ejemplos suministrados fuesen de cisnes blancos, es muy probable que la red no reconociera a un cisne negro como cisne.

Configuración inicial

La razón por la cual las redes de neuronas artificiales necesitan un entrenamiento tan preciso y extenso es que su conectoma de partida, es decir, su mapa de conexiones iniciales, es *aleatorio*, mientras que el conectoma del cerebro humano está parcialmente determinado, tal y como establece el *principio de especificidad de las conexiones* de Ramón y Cajal (Kandel, Schwartz & Jessell, 1995, p. 25). En el capítulo cuarto dijimos que, en el embrión, las neuronas del sistema nervioso

central se generan en las zonas ventriculares del neuroepitelo. Desde ahí migran hasta su destino sirviéndose, al parecer, de un mecanismo de afinidad química (Ibíd., p. 102). Según la hipótesis de la afinidad química, que es la más aceptada, las neuronas adquieren en un momento temprano de su desarrollo unos marcadores químicos que son afines a ciertas sustancias y repelen otras. Así es como cada neurona encontraría su destino. Obviamente, la determinación de las conexiones por este mecanismo no es exacta, neurona por neurona, pues, como señala Churchland, es matemáticamente imposible codificar en el DNA humano tanta información como la del emplazamiento particular de 10^{11} neuronas y 10^{14} sinapsis (Churchland, 1989, p. 285). Además, muchas de las neuronas y de las sinapsis generadas durante la gestación desaparecen al poco tiempo, y es necesario que así sea, pues de lo contrario se producen patologías como el autismo, caracterizado por un exceso de neuronas en la corteza prefrontal.

Por tanto, la estructura de las redes de neuronas determinada por factores genéticos contiene sólo un plan maestro de organización. Su origen es evolutivo. Los primeros peces y vertebrados, dotados de sistema nervioso, aparecieron hace 550 millones de años (Brooks, 1991, p. 396). Desde entonces, cada individuo que ha nacido con mutaciones exitosas para la supervivencia en la disposición inicial de sus redes de neuronas se ha reproducido más que el resto, y en consecuencia su mutación se ha propagado. En cambio, como señala Rumelhart, los sistemas conexionistas no tienen esa ventaja de experiencia acumulada a nivel filogenético, sino que todos ellos empiezan desde cero, con configuraciones aleatorias (Rumelhart, 1989, p. 231). La solución a este problema pasa por iniciar los sistemas en configuraciones ya conocidas que, de alguna manera, reflejen las habilidades innatas que tenemos los seres humanos. Hay tres formas de lograr esto: por ensayo y error, estudiando los mecanismos de afinidad química que intervienen en la neurogénesis y mediante la evolución artificial. Veámoslas en este orden. La primera vía, desde luego, si se pretende construir una auténtica IA fuerte, queda descartada por presentarse como una empresa titánica, pues sería necesario simular en el laboratorio un proceso de ensayo y error que a la naturaleza le ha llevado 550 millones de años.

En cuanto al estudio de la *afinidad química*, también es complejo, pero aún así es más razonable que el puro ensayo y error, pues se basa en el análisis del resultado de la evolución. Las primeras moléculas que sirven de guía a los axones para encontrar su destino se descubrieron en los años 70, y en la actualidad se conocen cinco familias de estas moléculas. Paradójicamente, el científico que descubrió el principio de especificidad de las conexiones fue un español, Santiago Ramón y Cajal, padre de la neurología moderna, pero los continuadores de su tarea en este ámbito no parece que vayan a ser también españoles. O quizás sí, pero lo harán en el extranjero. En el diario *El País* del 6 de febrero de 2012 se publicó una entrevista a la bióloga española Eloísa Herrera, experta en afinidad química formada en Granada, Madrid y la Universidad de Columbia de Nueva York. Su diagnóstico de la investigación en nuestro país es lapidario: «El principal cuello de botella viene cuando tienes treinta y largos años y quieres volver a tu país a aplicar lo que has aprendido. Incorporarte a un grupo ya existente en España no es fácil, porque hay muy pocos contratos para *posdoct* senior, y crear o consolidar aquí tu propio grupo de investigación hoy es casi misión imposible: ni las universidades ni el CSIC sacan plazas. [...] Al final, lo invertido en la formación de investigadores en España revierte en otros países como EE.UU., Reino Unido o Alemania». Descubrir los mecanismos de afinidad química que participan en la neurogénesis es, en definitiva, crucial para la IA subsimbólica, y si algún día se logra, nos enteraremos leyendo la prensa extranjera.

Respecto a la *evolución artificial*, es una técnica que consiste, como dijimos en el capítulo anterior, en la «simulación en un ordenador del mismo procedimiento que ha tenido lugar a lo largo de millones de años en el mundo natural» (Santos & Duro, 2005, p. 37). De ese modo, en última instancia se obtendrían redes de neuronas artificiales con configuraciones iniciales semejantes a las determinadas en el cerebro humano por afinidad química durante la neurogénesis. Recordemos las dos grandes diferencias anteriormente señaladas entre la evolución natural y la artificial. La primera, que la evolución artificial es *dirigida*, mientras que la natural es ciega. La segunda, que la artificial procede *simulando* la evolución de individuos virtuales en

entornos también virtuales. Esto último supone un gran problema, ya que la calidad del individuo cuando sus estructuras de control sean implementadas en un robot dependerá en gran medida de la semejanza entre la simulación y el mundo real. Lo habitual, apuntan José Santos y Richard Duro, es que al realizar ese *salto a la realidad* se produzca una pérdida de calidad, es decir, que el individuo no se comporte con tanta destreza como la hacía cuando él mismo y su entorno eran simulaciones informáticas (Ibíd., p. 162).

Para evolucionar estructuras de control de aspiradoras autónomas, la simulación del entorno es factible, pues el mundo de una aspiradora se reduce a suelos, paredes y muebles que se puedan interponer en la trayectoria. La pérdida de calidad que sufrirán este tipo de seres artificiales al dar el salto a la realidad puede reducirse introduciendo *ruido* en las simulaciones. El ruido es un concepto que refiere a la irregularidad del mundo real: irregularidad tanto en el cuerpo del sujeto, que por ejemplo en el caso de la aspiradora se moverá con ruedas propulsadas por motores ligeramente distintos, como en el entorno, que nunca estará formado por superficies perfectamente lisas. Si, por el contrario, la simulación se realizase sin ruido, entonces se estaría recreando un micromundo, es decir, un modelo abstracto de la realidad. Mientras que la estrategia de la IA durante varias décadas, como ya hemos visto, fue crear individuos complejos de partida en los entornos simplificados de los micromundos, la robótica actual se basa en el enfoque diametralmente opuesto: empezar la tarea creando individuos simples capaces de habérselas con entornos tan complejos como la realidad misma (Ibíd., p. 36). De esta forma, se irían obteniendo individuos cada vez más complejos, siempre funcionales en el mundo real, hasta llegar a obtener seres humanos virtuales trasladables a robots.

Ahora bien, para obtener seres humanos virtuales es necesario simular en la computadora un mundo tan complejo como el real en el que se desarrollan los seres humanos. A diferencia de las aspiradoras, que tienen un mundo circundante (*Umwelt*) limitado a los suelos, las paredes y los muebles, el mundo humano es infinitamente complejo y lleno de ruido. A este respecto Santos y Duro reconocen que: «Modelar la

realidad con una precisión absoluta puede llevar a procesos de simulación computacionalmente intratables» (Ibíd., p. 162). O lo que es lo mismo: que simular el mundo real, como hacía una computadora en la película *The Matrix* (1999), aunque sólo sea en su dimensión puramente física, es imposible. Por tanto, lo más que se puede evolucionar artificialmente, por razones técnicas, son las estructuras de control de seres simples que se desenvolverán en entornos simples.

Otros problemas que limitan el alcance de la evolución artificial son los siguientes tres señalados por James Stone, de la Universidad de Sheffield: no existe una aproximación paradigmática, los agentes artificiales carecen de muchos tipos de aprendizaje, y las técnicas actuales sólo modelan un conjunto finito y pequeño de las propiedades de los sistemas evolutivos naturales (Ibíd., p. 185). Primero, que no existe una aproximación paradigmática universalmente aceptada apunta al hecho, expuesto en el capítulo anterior, de que hay numerosas arquitecturas de interconexión posibles, multitud de operadores de selección, y no menos formas de realizar la evolución, como por ejemplo con o sin herencia lamarquiana. Segundo, que los agentes artificiales no poseen muchos tipos de aprendizaje señala que los individuos simulados son capaces de aprender por condicionamiento clásico, condicionamiento operante, o aprendizaje vicario, pero no de otras maneras más sofisticadas de aprendizaje social, como por ejemplo el lenguaje. Y tercero, las técnicas actuales sólo simulan un conjunto finito y pequeño de las propiedades de los seres reales porque la simulación total de, pongamos por caso, todos los procesos que suceden en una red de neuronas natural, es a día de hoy imposible. Ya aportamos el dato de que, según Henry Markram, director del proyecto Blue Gene, para simular el cerebro humano tal y como lo conocemos haría falta una máquina «20.000 veces más potente que los actuales superordenadores, con una memoria capaz de almacenar 500 veces todo el contenido del sistema actual de Internet» (Kaku, 2011, p. 138).

Por estas razones, y por muchas otras que no recogemos para no extendernos demasiado, la evolución artificial, a pesar su veloz desarrollo en algo más de dos décadas de andadura, no se presenta como una técnica capaz de descubrir las

configuraciones iniciales del cerebro humano, con sus consiguientes disposiciones innatas para el lenguaje, el reconocimiento de objetos, y en general todas aquellas que nos permiten aprender de forma rápida y eficaz, a diferencia de las redes de neuronas artificiales, que por empezar su vida con configuraciones aleatorias necesitan de entrenamientos muy largos para compensar su falta de herencia evolutiva.

El cuerpo

En conclusión, la única forma de descubrir configuraciones iniciales de las redes de neuronas que reflejen nuestras disposiciones innatas heredadas de la especie parece ser el estudio de la afinidad química. A modo de estrategia argumentativa, supongamos que la doctora Herrera y sus colegas, después de muchos años estudiando la neurogénesis de las 300 neuronas del *caenorhabditis elegans* y de las 100.000 de la *drosophila melanogaster*, consiguieran su objetivo final: descubrir al detalle la neurogénesis del ser humano. Retomando la metáfora de las nuevas gafas de Kant que propusimos en el capítulo quinto, lo que habrían conseguido es una descripción de las lentes genéticas universales de nuestra especie. La aplicación inmediata para la IA subsimbólica sería que, copiando esas configuraciones iniciales, los sistemas conexionistas aprenderían tan rápido como nosotros. Ahora bien, habría que preguntarse entonces *cómo* aprenderían en el sentido de cuáles serían sus entradas sensoriales y cuáles los miembros movidos por sus motoneuronas, pues para la formación de un cerebro adulto, plenamente inteligente, tan necesario como el cerebro es el medio material de entradas y salidas de información: el cuerpo.

Pensar que todo se reduce al cerebro sería caer en una especie de cartesianismo materialista contemporáneo, en el que se sustituye el alma por el cerebro, pero el cuerpo sigue siendo prescindible en la explicación del pensamiento. La inmersión en el mundo físico y social, que en la sección anterior hemos reclamado como indispensable para la formación de la inteligencia, acontece a través del cuerpo, por lo que el modelado del cerebro a causa de la experiencia depende de cómo sea

percibida dicha experiencia a través del cuerpo. Sin embargo, la IA, históricamente, ha adoptado un enfoque cartesiano, incluso en el programa de investigación subsimbólico, en tanto que ha pretendido crear inteligencias artificiales sin cuerpo (Copeland, 2004, p. 439; Haugeland, 1996, p. 25). Es un enfoque que en capítulos anteriores hemos denominado *IA abstracta*, frente a la *IA situada* que sería la que reconoce la importancia del cuerpo, como hace la robótica, y por tanto se esfuerza en replicarlo. En este apartado vamos a argumentar por qué el enfoque tradicional de la IA abstracta es erróneo. Antonio Damasio tiene razón cuando afirma que: «La comprensión global de la mente humana [...] debe relacionarse con un organismo completo, formado por la integración del cuerpo propiamente dicho y el cerebro, completamente interactivo con un ambiente físico y social» (Damasio, 1994, p. 288).

Pensemos en la oración "Juan disfrutó ayer montando en bicicleta por el parque". Una sentencia de este tipo, dice Dreyfus, apela a nuestra capacidad para imaginar cómo nos sentiríamos nosotros en esa situación, y no a la consulta de hechos relativos a bicicletas y parques, o a cómo reaccionaría un ser humano (Dreyfus, 1992, p. xix). El significado, por tanto, refiere a nuestra experiencia montando en bicicleta. Una IA fuerte, ya sea simbólica o subsimbólica, desprovista de cuerpo, no podría experimentar la sensación de montar en bicicleta, y en consecuencia el significado que le daría a la oración no se parecería al que le daría un ser humano que sí la haya experimentado. En este punto surgen dos observaciones. La primera, que hay personas que no saben o no pueden montar en bicicleta y, sin embargo, hasta cierto punto saben lo que significa, en tanto que en una conversación sobre el tema saben utilizar esa palabra. La segunda, que la experiencia de montar en bicicleta podría ser sustituida por una simulación virtual.

Empezando por esta última, ciertamente se puede programar en una computadora un simulador de ciclismo en primera persona, como se pueden programar simuladores de pilotaje de aviones y de muchas otras experiencias. Ahora bien, las simulaciones tienen un límite. Como acabamos de señalar, es imposible crear una simulación informática del mundo en su totalidad al estilo de *The Matrix*, y no sólo

por las mencionadas razones técnicas de carga computacional. Si semejante proyecto se acometiera con un planteamiento no evolutivo para acelerar el proceso, entonces requeriría simular seres humanos virtuales que constituyesen la dimensión social de mundo. Pero esos seres humanos virtuales son justo lo que se pretende obtener, por lo que se caería en un bucle en el que la simulación necesitaría a su vez de otra simulación del mundo que permitiera a esos seres humanos virtuales adquirir las experiencias necesarias para llegar a ser inteligentes.

La otra objeción es más seria. Hay personas que no saben o no pueden montar en bicicleta y sin embargo saben, *hasta cierto punto*, lo que es montar en bicicleta, en tanto que en la conversación demuestran saber jugar a los juegos de lenguaje en los que participa la expresión "montar en bicicleta". Es importante subrayar la cláusula "hasta cierto punto". Imaginemos a alguien que nunca haya montado en bicicleta, ni como piloto ni como pasajero. El conocimiento de alguien así sobre lo que significa "montar en bicicleta" se limitaría a las experiencias de haber visto cómo otras personas lo hacen y a las experiencias transmitidas por los relatos sobre lo que se siente al hacerlo. Sería una situación similar a la de los sacerdotes de ciertas religiones, que imparten cursillos matrimoniales siendo célibes.

Por un lado están las experiencias de ver cómo otros montan en bicicleta, y por otro las experiencias a través de narraciones. En cuanto a las primeras, el sujeto puede compartirlas empáticamente porque sabe lo que es tener un cuerpo. Ve al ciclista inclinarse en las curvas, y él sabe lo que se siente al estar inclinado. Ve al ciclista dar botes por los baches del camino, y él sabe lo que se siente al atravesar una superficie irregular. Respecto a las experiencias a través de narraciones, sucede lo mismo. Cuando escucha que en bicicleta se experimenta una gran sensación de velocidad, sabe lo que es la velocidad por haber viajado en otros medios de transporte. Cuando decimos que su comprensión de lo que es montar en bicicleta llega "hasta cierto punto", lo que queremos indicar es que hay experiencias que superan su imaginación porque no hay ningún paralelismo entre algunas de las sensaciones que se experimentan al montar en bicicleta y las que él haya podido experimentar de otras

formas. Por ejemplo, la combinación simultánea de estar inclinado mientras atraviesa un terreno accidentado a toda velocidad. En cualquier caso, incluso suponiendo que la comprensión indirecta del significado a través de la observación y del lenguaje fuera indistinguible de la comprensión directa por experimentación, la indirecta se basa siempre en la elicitación de memorias de otras experiencias vividas con el cuerpo, por lo que en última instancia es necesario un cuerpo. El significado de "velocidad", por ejemplo, es un patrón de activación de las neuronas del cerebro que sólo se puede obtener experimentando la velocidad a través del cuerpo.

Por tanto, a menos que se quiera sostener un dualismo de sustancias, el cuerpo es tan necesario para la inteligencia como el cerebro. Para superar el test de Turing es necesario tener cuerpo o, como diría Philip K. Dick, tener al menos el *recuerdo* de haberlo tenido. Y el cuerpo de una inteligencia artificial, dejando al margen especulaciones de ciencia ficción, no parece que pueda ser otra cosa si no un cuerpo artificial, es decir, un robot. La *Asociación de la Industria Robótica Americana* (*RIA*) define *robot* como «un manipulador reprogramable y multifuncional diseñado para mover material, partes, herramientas, o dispositivos especializados mediante movimientos variables programados para la realización de una variedad de tareas» (Santos & Duro, 2005, p. 4). Este uso fue acuñado por vez primera por el escritor checo Karel Capek en su obra de teatro de 1921 *R.U.R.* (*Rossum's Universal Robots*). "Robot" deriva de la palabra eslava *robota*, que significa "trabajo" o "servidumbre".

En la actualidad los robots están presentes en casi todo tipo de tareas productivas, habiéndolos de muy distintos grados de complejidad. Entre los más simples, pero a la vez más provechosos para el progreso de la robótica, está el *Khepera*. El robot Khepera es el más utilizado en robótica evolutiva desde su creación a mediados de los 90 (Ibíd., p. 33), aunque en la actualidad compite con alternativas más modernas como el *Koala*. La morfología del Khepera se reduce a una forma cilíndrica, de unos 6 centímetros de altura, con sensores de luz e infrarrojos para detectar objetos y con dos pequeñas ruedas que le sirven de actuadores. Como la evolución artificial se suele realizar mediante simulación informática, el Khepera sólo se usa al

final o en contadas generaciones en medio del proceso evolutivo. Ahora bien, a pesar de la utilidad del Khepera y similares, se trata de cuerpos muy simples. Por lo general, cuanto más complejo es un robot, menos inteligente es. Este hecho es un problema importante para la IA fuerte, dado que el cuerpo humano es extraordinariamente complejo, repleto de sensores y actuadores compuestos por multitud de subsistemas. Sin embargo, como dice Descartes, no hay argumentos de principio que impidan la replicación perfecta del cuerpo humano, pues no es más que una máquina (Descartes, *Discurso del método*, p. 92).

En consecuencia, al no haber argumentos de principio contra la posibilidad técnica de la replicación del cuerpo humano ni contra la replicación de las redes de neuronas de nuestro cerebro, la IA subsimbólica en sentido fuerte es técnicamente posible. Mientras que la IA simbólica es imposible por principio por las razones epistemológicas expuestas en las secciones anteriores de este capítulo, la realización de la IA subsimbólica es una cuestión empírica cuya viabilidad depende de la azarosa inspiración de los científicos y del interés social que haya en ella.

Para qué

¿Merece la pena invertir cantidades ingentes de recursos humanos y materiales en darle vida a Galatea? En darle vida a Galatea, tal vez no, pero en darle vida a una IA como la temida por Weizenbaum, capaz de interceptar todas las telecomunicaciones mundiales y ofrecer resúmenes de las relevantes para ciertos propósitos, sí (Weizenbaum, 1976, p. 223). En darle vida a un golem físicamente superior a cualquier hombre, y sin ningún escrúpulo moral para cumplir su misión como hacía el robot de la película *Terminator* (1984), también. Y en darle vida a una IA del tipo *Deus ex machina*, como la del relato *Un conflicto evitable* de Asimov, capaz de calcular la economía mundial, sin duda que también (Asimov, 1950, p. 336). Prueba preliminar de estas afirmaciones es que DARPA lleva medio siglo invirtiendo en el desarrollo de la IA a pesar de los continuos fracasos. La Historia de la IA está ligada a las organizaciones

militares como DARPA (Crevier, 1993, p. 313), un hecho que infunde el temor fundado de que la IA fuerte, si alguna vez se logra, no se utilizará, citando a Herbert Marcuse, para la pacificación de la existencia (Marcuse, 1964, p. 41). Sobre este tipo de cuestiones, relativas a las condiciones de posibilidad sociales de la IA fuerte, vamos a ocuparnos en el siguiente y último capítulo, el octavo.

8. Condiciones de posibilidad sociales

Para el desarrollo de una tecnología no sólo es necesario que ésta sea posible desde el punto de vista técnico, sino también desde el punto de vista social, es decir, que sirva a los intereses de, por lo menos, una parte de la sociedad. Así, en el capítulo segundo vimos que la notación indoarábiga, introducida en Europa a nivel académico por Fibonacci a comienzos del siglo XIII, no fue adoptada por los comerciantes hasta tres siglos después, debido a que éstos no la encontraban útil por razones tales como que las cantidades eran fácilmente falsificables añadiendo ceros a la derecha o que requería del uso de papel, un artículo de lujo en la época (Guijarro & González, 2010, p. 51). De la misma manera, tenemos que preguntarnos si la IA fuerte ofrece alguna utilidad a aquellos que pueden costearla. Dado que ser trata de una tecnología que, tal y como la hemos descrito en los dos capítulos anteriores, requiere de grandes inversiones, los únicos sujetos que podrían adquirirla en sus primeras versiones serían las corporaciones empresariales y los Estados.

El poder estatal moderno, decían Marx y Engels, no es más que «un comité que administra los negocios comunes de la clase burguesa, globalmente considerada» (Marx & Engels, *Manifiesto del partido comunista*, p. 50). Casi dos siglos después, esta sentencia sigue siendo verdadera, con la única puntualización de que la clase burguesa actual no es como la del siglo XIX. Como dice Horkheimer: «En la actual época de la gran industria el empresario independiente ya no es una figura típica» (Horkheimer, 1947, p. 152). Su lugar en el mundo ha sido ocupado por los monopolios de las corporaciones empresariales. Así pues, se podría parafrasear a Marx y Engels afirmando que el poder estatal contemporáneo no es más que el comité que administra los negocios comunes de las corporaciones. Las democracias, que

supuestamente deberían utilizar el poder estatal para contrarrestar los desequilibrios generados en la sociedad civil por la libertad de la propiedad privada (Marcuse, 1941, p. 199), en realidad están en manos de políticos cuyas campañas propagandísticas, condición necesaria y en muchos casos suficiente en la era de la cultura de masas para llegar al poder, son sufragadas por las corporaciones, razón por la cual a ellas se deben cuando gobiernan, y en sus consejos de administración terminan recalando como pago por los servicios prestados. Son los intereses de las corporaciones, por tanto, los únicos que debemos descubrir para determinar la posibilidad social de la IA fuerte.

Dividiremos este capítulo en tres secciones. En la primera analizaremos el perfil psicológico de las corporaciones, cosa que haremos de la mano de Robert Hare, el mayor especialista mundial en psicopatía. En la segunda utilizaremos la crítica de la razón instrumental de Horkheimer y de otros filósofos de la escuela de Frankfurt para desvelar las razones históricas por las que dicho perfil ha llegado a ser el dominante. Y en la tercera resolveremos, en primer lugar, cuáles son los intereses hacia a la IA fuerte que tiene la sociedad de nuestro tiempo, sometida al mandato de las corporaciones y, a continuación, examinaremos la posibilidad de que la IA fuerte pueda, por sí sola, producir un cambio cualitativo de la sociedad.

8.1. Una sociedad enferma

«Si no pudiésemos estudiar a los sociópatas en la cárcel, mi siguiente elección sería probablemente la bolsa de Vancouver» (Hare, 1993, p. 158). Éste es el comentario jocoso pero realista del doctor Hare sobre un artículo de Forbes en el que se describía la bolsa de Vancouver como un lugar infestado de personas deshonestas. Hare enumera tres razones por las cuales los psicópatas se sienten sumamente atraídos por la delincuencia de cuello blanco, la cual alcanza su niveles más altos en las corporaciones, ya sean éstas de producción de bienes materiales o financieros, como las que se dedican a mover el dinero en el mercado de valores. La primera razón es que son entornos que presentan multitud de oportunidades para conseguir lo que los

psicópatas más desean: el poder y el control sobre los demás (Ibíd., p. 62). La segunda es que los psicópatas tienen todas las cualidades necesarias para estafar a los demás: «facilidad de palabra, encanto, seguridad en sí mismos, control sobre las situaciones sociales, frialdad bajo presión, no les asusta la posibilidad de que les pillen y no tienen piedad» (Ibíd., p. 160). Y la tercera es que el delito de cuello blanco ofrece grandes recompensas a cambio de un riesgo mínimo de ser castigado. Y cuando se ha de afrontar el castigo, éste suele consistir en sanciones triviales gracias a que, como acabamos de señalar, los que realizan este tipo de actos «juegan un papel decisivo en el establecimiento de las reglas de control gubernamental» (Ibíd., p. 160).

Nuestra tesis es que las corporaciones tienen el perfil psicológico de un psicópata o, dicho con más precisión, de un sociópata. La idea no es nuestra, sino que la debemos al documental *The corporation* (2003), de Mark Achbar y Jennifer Abbott. No obstante, vamos a ir más lejos que Achbar y Abbott, en tanto que ellos se limitan a atribuir dicho perfil sólo a las corporaciones, mientras que nosotros vamos a argumentar que la psicopatía no es un perfil psicológico que surja de las corporaciones por su estructura legal en contra de la voluntad de los individuos que las integran, sino que es la clave del éxito, en general, en la sociedad actual debido a las razones que veremos en la sección siguiente al exponer la crítica de la razón instrumental, y por tanto también es característico de las personas físicas que dirigen las personas jurídicas de las corporaciones. Una segunda diferencia entre nuestro discurso y el de Achbar y Abbott es que ellos califican de *psicópatas*, sin más, a las corporaciones, cuando lo correcto sería calificarlas de *sociópatas*. Según el doctor Hare, la psicopatía es un *síndrome*, es decir, un conjunto de síntomas relacionados (Ibíd., p. 57). Ante un paciente que mostrase la mayoría de esos síntomas, el diagnóstico arrojado por dos especialistas en la materia podría ser tan divergente como que uno lo etiquetase de psicópata, y el otro, de sociópata (Ibíd., p. 44). La diferencia entre ambas patologías, dice Hare, es que la sociopatía está causada por factores exclusivamente sociales, mientras que a la psicopatía contribuyen además factores biológicos. Dado que las corporaciones no tienen biología, sería más acertado calificarlas de sociópatas.

Para terminar de aclarar la terminología, Hare señala un tercer trastorno que a menudo se confunde con la psicopatía y la sociopatía: el *trastorno de personalidad antisocial*. A grandes rasgos, la diferencia consiste en que la mayoría de los criminales cumplen los criterios para el diagnóstico de personalidad antisocial, mientras que la psicopatía es mucho menos frecuente, y se define «por un conjunto de rasgos de la personalidad y conductas socialmente desviadas» (Ibíd., p. 45). Dado que nuestro interés no es investigar estas cuestiones clínicas, y dado que vamos a utilizar como referencia a Robert Hare, quien se declara partidario de denominar al síndrome que nos ocupa como psicopatía en lugar de sociopatía, en esta sección vamos a trazar el *perfil psicológico* de las corporaciones y, en general, de los triunfadores que dirigen los destinos de la sociedad actual, refiriéndonos a él como psicopatía. En sentido estricto, las corporaciones no pueden ser psicópatas porque carecen de biología, y la mayoría de los individuos que las dirigen no son psicópatas en tanto que su conducta no tiene base biológica, sino que es aprendida, pero denominaremos psicopatía a su síndrome por darle el mismo nombre que le da Hare y así simplificar la exposición en aras de una mayor claridad. Lo que aquí nos importa son los *síntomas* del síndrome, es decir, los rasgos de personalidad de aquellos que dirigen el mundo mediante los monopolios, para determinar después los intereses que personalidades de ese tipo, que son las que disponen de los recursos económicos necesarios para desarrollar la IA fuerte, pueden tener respecto a esta tecnología.

Psicopatía

Para diagnosticar la psicopatía Hare y sus colaboradores desarrollaron en los años 80 el *Psychopathy Checklist*, un test que, como él apunta, proporciona también «un retrato detallado de la personalidad de los psicópatas» (Ibíd., p. 55). Los síntomas clave de los que se compone el síndrome de la psicopatía son de dos tipos o factores: emocionales y de desviación social. Los *emocionales* son seis: mente simple y superficial; personalidad egocéntrica y presuntuosa; falta de remordimientos o culpa;

falta de empatía; persona manipuladora y mentirosa; y portador de emociones superficiales. Por su parte, los de *desviación social* son otros seis: impulsividad; poco control de la conducta; necesidad de excitación; falta de responsabilidad; problemas de conducta en la infancia; y conducta antisocial de adulto. Este listado de síntomas no es tan exhaustivo como el de la última versión del *Psychopathy Checklist*, denominada PCL-R, pero es válido para nuestros intereses, que son filosóficos, y no clínicos.

Con el propósito de elucidar si las corporaciones tienen, en efecto, el perfil de un psicópata, Achbar y Abbott dedican una sección de *The Corporation* a enumerar una serie de comportamientos de las corporaciones que han trascendido a la prensa durante las últimas décadas. A continuación, se los presentan al propio Hare, quien sentencia ante las cámaras que alguien que se comportase de tal manera, si fuera una persona física, sería un psicópata. El diagnóstico de la psicopatía, dice Hare, «se lleva a cabo *sólo* cuando hay una evidencia sólida de que un individuo tiene el perfil completo de psicópata, esto es, cuando presenta la mayoría de los síntomas» (Ibíd., p. 97). Según él, a la corporación «le serían aplicables todos los síntomas y, de hecho, en muchos sentidos, la corporación tiene el perfil prototípico del psicópata». Veamos con más detalle los rasgos de la personalidad psicopática de las corporaciones.

Hay que comenzar advirtiendo que, según los cánones legales y psiquiátricos, los psicópatas no están locos, dice Hare (Ibíd., p. 25). A diferencia, por ejemplo, de los psicóticos, los psicópatas no experimentan alucinaciones o delirios que les impidan percibir correctamente la realidad. Son conscientes de lo que hacen y por qué lo hacen. «Su conducta es el resultado de una elección libremente ejercida» (Ibíd., p. 42). Dicho esto, el primer síntoma emocional es la *mente simple y superficial*. Esto implica que no les importa que les descubran mintiendo, algo que hacen de forma constante para conseguir sus propósitos. Y si se descubre la mentira, les trae sin cuidado, ya que tienen la habilidad característica de «contradecirse a sí mismos de una frase a la siguiente» (Ibíd., p. 63). El segundo síntoma es la *personalidad egocéntrica y presuntuosa*, lo que significa, dice Hare, que los psicópatas tienen una visión narcisista de la vida. «Se creen el centro del universo, seres superiores a los que se debiera

permitir vivir según sus propias reglas» (Ibíd., p. 61). El mundo entero sólo existe para satisfacer sus deseos. El tercer síntoma es la *falta de remordimientos o culpa*. Nunca se arrepienten del dolor que han causado, y cuando declaran arrepentimiento, lo hacen siempre en falso, pues sus actos desmienten sus palabras. Su notoria habilidad para distorsionar la verdad les sirve incluso para presentarse ellos como víctimas cuando en realidad son los verdugos (Ibíd., p. 52). El cuarto síntoma, la total *falta de empatía*, es, a juicio de Hare, el factor fundamental implicado en la psicopatía (Ibíd., p. 25). Los psicópatas tienen teoría de la mente, lo que significa que son capaces de ponerse en el lugar del otro en el sentido de hacerse una imagen de lo que está pensando, una habilidad imprescindible para manipular, como ellos hacen, las voluntades ajenas. Pero los sentimientos de los demás no son de su interés. Para ilustrar este rasgo de personalidad Hare cita al psicólogo Robert Rieber: «En el mundo del psicópata no existe el meramente débil. El que es débil también es un imbécil, esto es, alguien que pide que le exploten» (Ibíd., p. 69). Según Hare, las afirmaciones de los psicópatas revelan «su creencia de que el mundo se divide en "los que dan y los que toman", depredadores y víctimas, y que serían muy tontos si no explotasen las debilidades de los demás» (Ibíd., p. 74). El quinto síntoma es la *personalidad manipuladora y mentirosa*, de la que venimos hablando desde el principio. Y el sexto es que el psicópata es *portador de emociones superficiales*. Sólo poseen emociones básicas e incompletas, como se desprende del hecho de que ante el miedo no presentan las respuestas fisiológicas normales del resto de la población (Ibíd., p. 80).

En cuanto a los síntomas de desviación social, el primero es la *impulsividad*. Los psicópatas viven en el presente, sin darle mucha importancia al pasado y el futuro. El tiempo para ellos es casi diacrónico más que dialógico, atendiendo a la distinción de Heidegger que expusimos en el capítulo quinto. Cuando tienen un deseo, simplemente lo satisfacen, ignorando las consecuencias y las necesidades de los demás. El segundo es el *poco control de la conducta*, lo que implica que no reprimen sus reacciones más coléricas. Si piensan que alguien es un obstáculo para ellos, lo eliminan sin más consideraciones. El tercero es la *necesidad de excitación*. Los psicópatas necesitan vivir

al límite, un rasgo derivado quizás de la escasa profundidad de sus emociones. Siempre quieren más de lo que tienen. En su caso se cumple con particular perfección la definición del deseo formulada por Quevedo: el deseo es aquello que hace que lo mucho se vuelva poco con sólo desear un poco más. El cuarto síntoma es la *falta de responsabilidad*. Las obligaciones y los compromisos, dice Hare, no significan nada para los psicópatas (Ibíd., p. 89). Prometen a menudo, pero sólo como estrategia para lograr sus fines, no porque vayan a cumplir lo pactado. El quinto síntoma es que suelen mostrar *problemas de conducta en la infancia*, que van desde el robo hasta el vandalismo, pasando por las amenazas y el chantaje. Finalmente, el sexto síntoma es la *conducta antisocial del adulto*. «Los psicópatas consideran que las reglas y expectativas de la sociedad son sólo inconvenientes, impedimentos poco razonables a la plena expresión de sus inclinaciones y deseos» (Ibíd., p. 95).

Para la demostración de que las corporaciones cumplen con la mayoría de estos síntomas, nos remitimos al citado documental de Achbar y Abbott, o bien, directamente a la lectura de los periódicos. Por mencionar un caso significativo, remontémonos a 2010. Apple, una corporación tan grande que su valor en bolsa supera al producto interior bruto de países como Argentina o Grecia, había contratado a Foxconn para que fabricase en China el iPhone 4. Cómo serían las condiciones de explotación laboral, que en los cinco primeros meses de ese año se suicidaron diez trabajadores. La solución fue colocar redes antisuicidio alrededor del edificio. De esa manera, los trabajadores dejarían de tirarse por la ventana. Podrían seguir acabando con sus vidas cortándose las venas en los lavabos, pero no arrojándose al exterior. Así se neutralizaba el riesgo de que un cuerpo esparcido sobre la acera fuera fotografiado. Apple quiere sus teléfonos a tiempo, y para eso, como haría un psicópata, no duda en explotar literalmente hasta la muerte a otros seres humanos. No aprovecharse de ellos sería estúpido, porque son débiles, y los débiles, como dice Rieber, son a ojos del psicópata imbéciles que merecen ser explotados. La producción, por supuesto, siguió adelante. Se cumplieron los plazos de entrega, que es lo único que importa a una personalidad narcisista, y los consumidores disfrutaron de sus iPhone 4.

Respecto a quienes dirigen las corporaciones, afirmamos que su personalidad también tiene muchos de los síntomas característicos de la psicopatía. Achbar y Abbott, en cambio, los retratan como simples trabajadores que se ven obligados a tomar decisiones inmorales en contra de su voluntad y de sus principios. Algo así como los oficiales nazis de los campos de concentración que después alegarían en los Juicios de Núremberg que ellos se limitaban a cumplir órdenes. Las órdenes de los directivos de las corporaciones son obtener beneficios en un mercado muy competitivo, y ellos cumplen con su deber. Al final, por tanto, la culpa es del mercado, es decir, de nadie en particular que pueda ser castigado. Es la repetición del pecado original: Dios pregunta a Adán por qué comió la fruta prohibida, él responde que la culpa es de Eva, ella le echa la culpa a la serpiente, y la serpiente desaparece reptando entre unos matorrales. La culpa se ha volatilizado. Así, nadie es responsable de que los trabajadores de Foxconn se suiciden. Nuestro punto de vista es bien distinto.

Esquizofrenia

Achbar y Abbott sugieren que las corporaciones son leviatanes que, como el Leviatán que describe Hobbes a propósito del Estado, son personas distintas de aquellas que los dirigen. Sin embargo, la realidad es que, como decía Ambrose Bierce, la persona jurídica no es más que un artificio para obtener ganancia individual sin responsabilidad individual. Detrás de las personas jurídicas siempre hay personas físicas que toman las decisiones y que se benefician de ellas. Son, por tanto, los directivos de las corporaciones quienes se comportan como psicópatas. Cuando los miembros del consejo de administración de Apple aprueban encargar la fabricación del iPhone 4 a Foxconn, con sede en China, saben que el beneficio económico derivado de esa decisión será a costa de la explotación de seres humanos en el tercer mundo. Pero ellos, en lugar de dimitir, toman la decisión de quedarse en la empresa y cobrar el *bonus*, igual que los sujetos experimentales de Milgram decidían pulsar el botón y electrocutar a personas inocentes (Worchel, Cooper, Goethals & Olson, 2000, p. 355).

En la sociedad industrial avanzada, por tanto, los grandes triunfadores, los que dirigen el mundo desde despachos parapetados tras las leyes que ellos mismos dictan, se comportan como psicópatas. Ésta es la norma de conducta de nuestro tiempo. El resto de la población, los que en lugar de un Rolls Royce conducen un utilitario o viajan en autobús, no dudan en aplicarla ellos también para triunfar en su microcosmos. Es un hecho que se refleja en el siguiente artículo de *The New York Times* citado por Hare: «El joven criminal que vemos hoy está más desvinculado de su víctima que nunca, más preparado para herir o matar. La falta de empatía por sus víctimas que vemos entre los jóvenes criminales es sólo un síntoma de un problema que afecta a toda la sociedad. La postura del psicópata es más común en la actualidad que antes; la sensación de ser responsable del bienestar de los demás disminuye día a día» (Ibíd., p. 111). Acerca de la fascinación pública en los últimos tiempos por las novelas y películas sobre psicópatas, el psiquiatra forense Ronald Markman tiene una teoría: el público se identifica con los psicópatas (Ibíd., p. 109). Gracias a esas historias, dice, liberamos nuestras fantasías de una vida sin controles internos. Hare se pregunta, escandalizado, si estamos permitiendo inconscientemente que la sociedad evolucione hacia convertirse en un "criadero de asesinos" (Ibíd., p. 111).

En efecto, la conducta psicopática se extiende porque es promocionada por la sociedad. Pero no abiertamente, claro está, sino a través de la *doctrina de la doble verdad*. Horkheimer lo ilustra con el caso del niño que «se da cuenta de que Papá Noel es, en realidad, un empleado de la tienda y que percibe la relación entre las Navidades y la cifra de ventas» (Horkheimer, 1947, p. 72). Por un lado, la sociedad entera, incluidos los grupos dominantes (Ibíd., p. 181), defiende una serie de valores morales y espirituales tales como los implicados en la celebración de la Navidad, mientras que, por el otro, se contradice a sí misma actuando en contra de dichos valores. Es lo mismo que hacen los psicópatas: decir una cosa y hacer la contraria. Este tipo de pluralismo, dice Horkheimer, introduce en la vida moderna un rasgo *esquizofrénico* (Ibíd., p. 73). Desde la adolescencia, el individuo percibe «el abismo que se abre entre los ideales en los que fue educado y el principio de realidad al que se ve obligado a someterse»

(Ibíd., p. 130). Le dicen que debe ser altruista, mientras que los triunfadores son egocéntricos; le dicen que debe ponerse en el lugar de los demás, mientras que los triunfadores carecen de empatía; le dicen que debe arrepentirse y pedir perdón por sus malas acciones, mientras que los triunfadores jamás cumplen sus propósitos de enmienda; le dicen que debe afrontar las consecuencias de sus actos, mientras que los triunfadores eluden la acción de la justicia; le dicen que no debe mentir, mientras que los triunfadores lo hacen por sistema. El individuo es, en definitiva, presa de una sociedad esquizofrénica que le obliga al *doblepensar* orwelliano: «Saber y no saber, hallarse consciente de lo que es realmente verdad mientras se dicen mentiras cuidadosamente elaboradas, sostener simultáneamente dos opiniones sabiendo que son contradictorias y creer sin embargo en ambas; emplear la lógica contra la lógica, repudiar la moralidad mientras se recurre a ella» (Orwell, 1949, p. 45).

Las respuestas posibles del individuo son dos. Por un lado, la *resistencia*. Este tipo de individuo, dice Horkheimer, es escaso. «En lugar de sacrificar la verdad conformándose y adaptándose a las pautas dominantes, intentará expresar en su vida tanta verdad como le resulte posible, tanto en la teoría como en la práctica. Llevará una vida conflictiva; tendrá que estar preparado para asumir el riesgo de una extrema soledad» (Horkheimer, 1947, p. 130). La otra respuesta, que es la de la mayoría, es la *sumisión*: «los que son demasiado débiles como para enfrentarse con la realidad no tienen más remedio que disolverse identificándose con ella» (Ibíd., p. 131). Esto es, la mayoría adopta la norma de conducta de la sociedad: la de la personalidad del psicópata, sólo que sin la sinceridad de éste. Mientras que el psicópata no tiene reparo alguno en reconocer abiertamente que su objetivo es el poder y la satisfacción de sus deseos pasando por encima de los demás, el individuo promedio de la sociedad actual lo hace de una manera esquizofrénica. El niño que descubre que Papá Noel es en realidad un empleado del comercio, dice Horkheimer, aprende pronto que no debe ser un aguafiestas (Ibíd., p. 73). Cuando el adulto lee en el periódico que el iPhone del que tanto presume es resultado de la explotación hasta el suicidio de seres humanos, lo que debe hacer es *doblepensar* la noticia y no ser un aguafiestas.

A propósito del artículo de *The New York Times* antes citado, al mostrar su preocupación por el auge de la conducta psicopática entre los más jóvenes, Hare se revela como lo que es: un hombre con gran sensibilidad social, pero con una mirada de alcance limitado por su disciplina, la psicología. La filosofía, en cambio, sabe que este patrón de conducta, caracterizado por la habilidad para conseguir fines sin someterse a restricciones morales, viene de muy atrás. Horkheimer y Adorno lo encuentran en plena Ilustración, a finales del siglo XVIII, en el retrato que hace el Marqués de Sade de la cara oculta de su época en obras como *Historia de Juliette*, *Filosofía en el tocador* y *Las ciento veinte jornadas de Sodoma* (Horkheimer & Adorno, 1944, p. 134). Y todavía se remonta mucho más atrás. Según Horkheimer y Adorno, a los orígenes mismos de la civilización. Esto es lo que significa su célebre sentencia: «El mito es ya Ilustración; la Ilustración recae en mitología» (Ibíd., p. 56).

8.2. La autoliquidación de la razón

En términos de la crítica de la razón instrumental de Horkheimer, el perfil del psicópata es el del sujeto dotado de razón subjetiva y carente de razón objetiva. La *razón subjetiva* o *instrumental*, a la que ya nos referimos en el primer capítulo, es definida por Horkheimer como una razón que «tiene que ver esencialmente con medios y fines, con la adecuación de los métodos y modos de proceder a los fines» (Horkheimer, 1947, p. 45). En cuanto a los fines, la razón subjetiva no suele ocuparse de ellos, y cuando lo hace, señala como racionales sólo los que sirven al sujeto para su autoconservación. La *autoconservación*, que no es otra cosa que el esfuerzo o *conatus* por perseverar en el propio ser, era considerada por Spinoza como la esencia actual de cada cosa (Spinoza, *Ética*, p. 191) y el primer y único fundamento de la virtud (Ibíd., p. 111). Esta última afirmación, dicen Horkheimer y Adorno, «contiene la máxima verdadera de toda civilización occidental» (Horkheimer & Adorno, 1944, p. 82). Ésta es, por tanto, la razón que impera en nuestro tiempo: la razón subjetiva, calculadora de medios y con el único fin egoísta de la autoconservación.

Sin embargo, prosigue Horkheimer, durante largo tiempo dominó una visión de la razón muy diferente: la *razón objetiva* o *axiológica*. Ésta, al contrario que la subjetiva, ponía el énfasis más en los fines que en los medios. «Grandes sistemas filosóficos, como los de Platón y Aristóteles, la Escolástica y el idealismo alemán, tenían como fundamento una teoría objetiva de la razón» (Horkheimer, 1947, p. 46). Dichos sistemas pretendían descubrir un orden objetivo en el universo, más allá de las apariencias empíricas siempre cambiantes, por lo que el grado de racionalidad de la vida de una persona podía medirse en función de su adecuación a dicho orden. La razón objetiva coexistía en aquel entonces con la subjetiva, pero subordinándola para poner sus cálculos de medios al servicio de fines más elevados que la mera autoconservación por la que se rigen las bestias.

Al orden objetivo del universo se ha accedido tradicionalmente mediante distintas vías. Hegel, máximo exponente del idealismo alemán, consideraba que la filosofía, el arte y la religión eran las maneras de descubrir el espíritu absoluto que debía guiar a la humanidad (Marcuse, 1941, p. 90). No obstante, esta pluralidad de caminos ha dado lugar a frecuentes conflictos, como prueba el registro histórico de las numerosas ocasiones en que la filosofía y la religión se han enfrentado. Por eso, la intención de la razón objetiva fue desde sus inicios «sustituir la religión tradicional por pensamiento filosófico metódico y por conocimiento e intelección» (Horkheimer, 1947, p. 52). El propósito de la razón objetiva era, por tanto, conferir un fundamento racional a las afirmaciones sobre el orden objetivo del universo. Lo que estaba en juego en la pugna entre filosofía y religión, dice Horkheimer, era el medio llamado a expresar la máxima verdad: revelación o razón, teología o filosofía.

Éste proceso histórico de avance de la razón es lo que se denomina, en sentido amplio, *Ilustración*. La Ilustración, dicen Horkheimer y Adorno, no se limita al lapso de tiempo que comienza en el Renacimiento y culmina en el Siglo de las Luces, sino que abarca el proceso histórico entero de racionalización. Este proceso «ha perseguido desde siempre el objetivo de liberar a los hombres del miedo y constituirlos en señores» (Horkheimer & Adorno, 1944, p. 59). La sentencia antes mencionada de que

«el mito es ya Ilustración» significa que el mito es Ilustración en tanto que es un instrumento del hombre para conocer la naturaleza y utilizar ese conocimiento para dominarla. «El mito quería narrar, nombrar, contar el origen: y con ello, por tanto, representar, fijar, explicar. [...] Los mitos, tal como los encontraron los Trágicos, se hallan ya bajo el signo de aquella disciplina y aquel poder que Bacon exalta como meta» (Ibíd., p. 63). La disciplina que Bacon exalta es la ciencia moderna, producto de la razón instrumental, y el poder que exalta como meta es la dominación para la autoconservación. Liberarse de la coerción de la naturaleza para imponer él sus propias reglas es lo que el hombre siempre ha deseado, y la Ilustración, que se inicia en el mito, es el proceso hacia ese fin.

Cierto es que la principal religión de Occidente, el cristianismo, erige al hombre en dueño y señor de la naturaleza, como se aprecia en la Biblia (Horkheimer, 1947, p. 124), pero lo hace a cambio de convertirlo en súbdito de Dios. La Ilustración, en tanto que proyecto de emancipación total, implica por tanto no sólo el sojuzgamiento de la naturaleza, sino también la liberación respecto de la religión. Y así lo hicieron los filósofos de la Ilustración: atacaron a la religión en nombre de la razón. Sin embargo, lo que hicieron fue no sólo derribar a la iglesia, sino que dieron la estocada definitiva «a la metafísica y al mismo concepto objetivo de razón, la fuente de poder de sus propios esfuerzos» (Ibíd., p. 56). De esta manera, en busca de la libertad y la autonomía, ha acontecido la paradoja de que «la razón se ha autoliquidado en cuanto medio de intelección ética, moral y religiosa» (Ibíd., p. 56).

Horkheimer y Adorno ponen a Kant como ejemplo de filósofo que intentó denodadamente salvar a la razón objetiva deduciendo el deber del respeto mutuo de una ley de la razón (Horkheimer & Adorno, 1944, p. 133). Sin embargo, esa ley carece de todo sostén crítico, pues no hay ningún argumento de la razón objetiva que justifique el imperativo categórico, como tampoco hay ningún argumento que obligue a la comunidad científica a cumplir los imperativos mertonianos. De la misma forma, tampoco hay ninguna demostración objetiva de que una sociedad orientada a ofrecer una mayor oportunidad para el libre desarrollo de las necesidades y las facultades

humanas es preferible a una sociedad represiva. Marcuse reconoce abiertamente que la elección de la primera se basa en un *juicio de valor*, y no en un juicio objetivo (Marcuse, 1964, p. 194). En consecuencia: «El burgués que se privara de una sola ganancia por el motivo kantiano del respeto a la mera forma de la ley no sería ilustrado, sino supersticioso: sería un loco» (Horkheimer & Adorno, 1944, p. 133).

Como dijimos en el capítulo quinto a propósito del chiste de Woody Allen sobre el hombre con un hermano que creía ser una gallina, la inteligencia consiste en ver lo que no está ahí, añadiéndolo el sujeto en base a la experiencia individual y de la especie. El loco y el genio se diferencian del resto en que ven lo que la mayoría no ve. Y la diferencia entre el loco y el genio radica, a su vez, en que las proyecciones del genio son socialmente valiosas, mientras que las del loco no lo son. En una sociedad guiada por la pura razón subjetiva, y que por tanto no cree en la razón objetiva, creer en el imperativo categórico es estar loco, porque es una visión supersticiosa, de algo cuya existencia no se puede probar empíricamente, y que además es irracional desde un punto de vista instrumental, pues aleja del éxito, es un obstáculo que impide la satisfacción del deseo. Hoy en día, por tanto, el loco es el kantiano y el cuerdo es el psicópata. Al desaparecer la razón objetiva, la razón subjetiva ha quedado libre de toda atadura: se erige en dueña y señora por encima de cualquier consideración moral.

En estos términos puramente instrumentales, la conducta psicopática es de todo punto racional, pues el psicópata es hábil como pocos manipulando a los demás para tomar de ellos lo que desea. Y, si no lo hiciera, sería un loco. Él es el verdadero sujeto cuerdo. Su perfil es el mismo que el de los personajes del Marqués de Sade: mienten, matan y explotan sin remordimiento alguno. Cuando le preguntaron por el sentimiento de culpa al famoso asesino en serie Ted Bundy, éste respondió con una frase que podría ponerse en boca de cualquiera de los cuatro libertinos de *Las ciento veinte jornadas de Sodoma*: «¿Culpabilidad? Eso es un mecanismo que usamos para controlar a la gente. No es más que una ilusión. Se trata de un mecanismo de control social, y es *muy* insano» (Hare, 1993, p. 65). Gracias a la falta de sentimiento de culpa, Ted Bundy pudo matar tranquilamente a doce jóvenes en los años 70.

Horkheimer y Adorno describen el comportamiento de los personajes del Marqués de Sade. Nosotros hemos descrito el de los psicópatas. Ambas formas son igualmente válidas para describir la catástrofe en que ha resultado la Ilustración. Esta catástrofe es el referente de la otra sentencia aforística de Horkheimer y Adorno: «la Ilustración recae en mitología». Parece contradictoria a la primera, pero no lo es. Para entender su significado, atendamos al cambio de la ciencia causado por la reducción de la razón a su dimensión subjetiva. Nada tiene que ver la ciencia anterior a Bacon y Galileo con la posterior a ellos. Cuando los hombres todavía se guiaban por la razón objetiva, la ciencia perseguía «descubrir una estructura omniabarcadora o fundamental del ser y derivar de ella una concepción del destino humano» (Horkheimer, 1947, p. 52). En cambio, una vez que se ha liberado de la razón objetiva, la ciencia moderna sólo aspira a entender la naturaleza, tanto interna como externa al sujeto, para dominarla. La reduce a mero sustrato de dominio, renunciando a entender el *sentido* que hay en ella (Horkheimer & Adorno, 1944, p. 61). Para comprender su sentido, dice Horkheimer, hay que «manejarla como un texto que ha de ser interpretado por la filosofía y que, leído correctamente, revela, en su despliegue, una historia de sufrimiento infinito» (Horkheimer, 1947, p. 141).

La manera en que el pensamiento interpreta esa historia y comprende su sentido es a través del concepto. El *concepto*, dice Marcuse, no designa simplemente la cosa particular, sino que abarca algo más: «alguna condición o relación universal que es esencial a la cosa particular, que determina la forma en que aparece como objeto concreto de la experiencia» (Marcuse, 1964, p. 108). Gracias a esta abstracción ampliadora de la mirada, que va más allá de lo concreto, el pensamiento auténtico relaciona los hechos con los factores que los provocan (Ibíd., p. 128). Los deshace para reconstruirlos en su verdadero ser, esto es, como parte de un proceso histórico. La democracia, por ejemplo, desde el punto de vista de la razón subjetiva se reduce a la descripción de aquello a lo que, en efecto, se denomina con ese nombre. Así es como piensan los pragmatistas como el segundo Wittgenstein: las palabras significan su uso, y nada más. En cambio, el concepto piensa la democracia sobre el horizonte histórico

de lo que ha sido y de lo que no es pero podría ser. Descubre los factores que han provocado que la democracia sea lo que es y denuncia las posibilidades que han sido reprimidas en el curso de la praxis: su enfoque es *molar*. Tal denuncia entraña un juicio de valor, naturalmente, que no puede ser demostrado por la ciencia moderna.

La ciencia moderna, en lugar de abstraerse de lo concreto para elevarse a lo universal, se abstrae de lo universal para centrar la mirada en lo concreto, cerrando la visión a todo lo demás. Opera con abstracciones, sin duda, pero esas abstracciones son mera síntesis técnica de datos fácticos (Horkheimer, 1947, p. 59) que presentan «lo universal sólo como la cara de lo particular por la que éste se deja captar y manipular» (Horkheimer & Adorno, 1944, p. 132). Su abstracción, por tanto, es reductora en vez de ampliadora: su enfoque es *molecular*. Sirve para manipular los hechos tal como se presentan, sin comprender las elecciones determinadas (Marcuse, 1964, p. 195) de la Historia de las que son consecuencia (Ibíd., p. 24). Los hechos, así tomados por sí solos, como átomos extraídos de su contexto, no tienen sentido. El análisis que de ellos se puede hacer «es "cerrado"; el campo de juicio se confina dentro de un contexto de hechos que excluye la posibilidad de juzgar el contexto en el que se forman los hechos, obra del hombre, y en el que su sentido, su función y su desarrollo están determinados» (Ibíd., p. 116). Marcuse ofrece también esta otra excelente definición del análisis funcional: «el análisis funcional se encierra en el sistema seleccionado que en sí mismo no es sujeto a un análisis crítico que trascienda las fronteras del sistema yendo hacia la continuidad histórica, en la que sus funciones y disfunciones llegan a ser lo que son» (Ibíd., p. 110). Recordemos la definición del comprender que dimos en el capítulo quinto: comprender es poner en relación con memorias pertinentes. La memoria es suprimida en la ciencia moderna, que, guiada por la sola razón subjetiva, en lugar de conceptos emplea fórmulas (Horkheimer & Adorno, 1944, p. 61).

La consecuencia es que, al no poder trascender los hechos en un sentido crítico y empírico (Marcuse, 1964, p. 21), el pensamiento se doblega a ellos: deviene así en positivismo y pragmatismo. «Lo que parece un triunfo de la racionalidad objetiva, la sumisión de todo lo que existe al formalismo lógico, es pagado mediante la dócil

sumisión de la razón a los datos inmediatos» (Horkheimer & Adorno, 1944, p. 80). Todo cuanto acontece ante una mirada de este tipo es pura repetición. La fórmula matemática nunca depara nada nuevo, sino siempre repetición de un hecho que, aislado del contexto, se presenta como igual. La caída en el *principio de la inmanencia*, según el cual nada hay nuevo bajo el sol (Ibíd., p. 67), es el precio que se paga. Esta repetición de todo cuanto ocurre, junto con la sumisión del pensamiento a la realidad dada, son características del mito. El mito está condenado a repetirse una y otra vez. Así se explica el significado de la segunda parte del aforismo: «la Ilustración recae en mitología». La liberación pretendida por la Ilustración ha resultado en un pensamiento, el de la razón subjetiva, que es esclavo de la realidad inmediata.

Esto es algo que hemos visto que también les sucede a los psicópatas: en tanto que olvidan su historia y viven en un eterno presente, son incapaces de gobernar su propia vida. «Los psicópatas tienden a vivir al día y a cambiar de planes frecuentemente. No le dan mucha importancia al futuro. No les preocupa lo que suceda mañana. De hecho, tampoco les importa mucho el pasado» (Hare, 1993, p. 85). En una ocasión, un psicópata le dijo a Hare: «Si pensase siempre en el mañana no sería capaz de vivir el presente» (Ibíd., p. 85). Gary Gilmore, famoso por haber solicitado al tribunal la pena de muerte para sí mismo, concedió una entrevista poco antes de su ejecución. El periodista le preguntó por qué le habían pillado tantas veces a pesar de su alto cociente intelectual, a lo que Gilmore respondió: «Nunca fui un gran ladrón. Soy demasiado impulsivo para ello. No planeo, no pienso. No tienes que ser superinteligente para salir ileso de toda esa mierda, sólo tienes que pensar un poco. Pero yo no pienso. Soy impaciente. Demasiado ávido. Podría haber escapado con un montón de golpes en mi haber. Pero no sé. Realmente, es difícil de entender. Quizá dejó de importarme hace mucho tiempo» (Ibíd., p. 119). Al olvidar su pasado, el psicópata está condenado a revivirlo una y otra vez, a entrar y salir de la cárcel por los mismos crímenes y los mismos errores. Cae en el principio de la inmanencia igual que ha caído la Ilustración por reducir el conocimiento al formalismo lógico. El formalismo no tiene memoria. Es una abstracción diacrónica ajena al tiempo dialógico.

La solución al fracaso de la Ilustración no es tema que nos concierna aquí, sino que para nuestros intereses es suficiente con describir el estado de cosas actual. No obstante, apuntemos que la solución, a juicio de Horkheimer, no pasa por retornar al dogma. No al dogma religioso, pero tampoco al de las metafísicas racionalistas, porque el paso de la razón objetiva a la subjetiva «no fue precisamente una casualidad y el proceso de evolución no puede reorientarse arbitrariamente en sentido contrario en un momento determinado. Si la razón subjetiva disolvió, bajo la forma de la Ilustración, la base filosófica de las convicciones fideístas que habían sido parte de la cultura occidental, ello fue posible porque esta base se había revelado como demasiado débil» (Horkheimer, 1947, p. 92). «Con ontologías reavivadas sólo se agrava la enfermedad» (Ibíd., p. 170). La "curación" pasa por identificar la causa del mal, que no es otra que el afán del hombre por dominar la naturaleza. Tal diagnóstico ha de ser realizado por la propia razón, por lo que la solución consiste, en definitiva, en ilustrar a la Ilustración: hacer reflexionar a la razón sobre sí misma. La filosofía, por tanto, debe renunciar a su tradicional pretensión del absoluto (Ibíd., p. 183) y adoptar un enfoque crítico, histórico y relativo. De esta forma, descubre que la causa que llevó a la formulación de los antiguos sistemas de la razón objetiva es que la razón subjetiva por sí sola es incapaz de garantizar la autoconservación supraindividual (Ibíd., p. 179).

Ambos conceptos de razón son necesarios. La deslegitimación positivista de la razón objetiva es tan perniciosa como el rechazo reaccionario de la razón subjetiva que carga la culpa sobre la técnica en sí misma (Ibíd., p. 162). El error radica en la hipóstasis de una sobre otra. La razón objetiva y subjetiva se necesitan recíprocamente tanto como el concepto y la intuición en el idealismo crítico de Kant, dice Horkheimer (Ibíd., p. 178). La primera proporciona el conocimiento teórico del *qué*, mientras que la segunda aporta el saber práctico del *cómo*. Este enfoque conciliador contrasta con las teorías de la inteligencia de Jeff Hawkins y Roger Schank que vimos en el capítulo quinto. Ambos identifican, si no totalmente por lo menos de manera prioritaria, la inteligencia con el saber práctico de la razón subjetiva, ignorando el teórico de la objetiva. Para Hawkins, recordemos, la inteligencia se mide por la capacidad de hacer

predicciones acertadas. Según sus propias palabras: «Conocer algo significa que puedes realizar predicciones al respecto» (Hawkins & Blakeslee, 2004, p. 125). Y otra sentencia suya aún más esclarecedora es la que dice que: «La inteligencia se mide por la capacidad predictiva de una memoria jerárquica, no por una conducta semejante a la humana» (Ibíd., p. 242). Schank, por su parte, revela un punto de vista similar al reclamar que la principal ventaja de la memoria es que proporciona el material necesario para realizar predicciones y generalizaciones (Schank, 1999, p. 55). O sea, lo mismo que hace la ciencia moderna, pero a escala del individuo. También es significativo el hecho de que cite a menudo a John Dewey como referente de su pensamiento (Ibíd., p. xi) y que proclame un concepto darwinista de inteligencia como habilidad procedimental para la autoconservación (Ibíd., p. 271).

Frente a estas teorías pragmáticas de la inteligencia que, por cierto, para ser coherentes deberían guardar silencio y abstenerse de teorizar (Horkheimer, 1947, p. 80), Horkheimer sostiene que: «Un hombre inteligente no es el que es simplemente capaz de hacer inferencias correctas, sino aquel cuyo espíritu está abierto a la percepción de contenidos objetivos» (Ibíd., p. 85). Nosotros estamos de acuerdo en este punto, razón por la cual de entre las teorías contemporáneas de la inteligencia defendimos como propia la de Howard Gardner, que en su vertiente práctica del Proyecto Zero está orientada al desarrollo pleno de las potencialidades del individuo desde la infancia, en lugar de atender exclusivamente al cultivo de las habilidades productivas demandadas por la economía. Al llegar a la edad adulta, los individuos así formados, por una pedagogía basada en la teoría de las inteligencias múltiples, serán inteligentes en el sentido que señala Horkheimer: capaces de percibir contenidos objetivos. Pero no de una manera dogmática, sino siempre crítica, reflexiva.

Sin embargo, la sola reflexión no es suficiente. En contra de las teorías de la resignación que, como el estoicismo, proponen la búsqueda de la realización interior mientras el mundo alrededor se hunde, Horkheimer sostiene que, para salvarse, además de reflexionar el individuo está obligado a salvar a la sociedad. No porque deba hacerlo por imperativo ético, sino porque ontológicamente no puede no hacerlo,

ya que «el individuo absolutamente aislado ha sido siempre una ilusión» (Ibíd., p. 148). El verdadero ser humano sólo existe en sociedad. La autosalvación, por tanto, implica salvar a la circunstancia. O, como decía Ortega: «Yo soy yo y mi circunstancia, y si no la salvo a ella no me salvo yo» (Ortega y Gasset, 1914, p. 35). Horkheimer, más optimista que Adorno, sí creía que la salvación es posible, aunque de una manera un tanto contradictoria, como veremos al final de la siguiente sección.

8.3. Un nuevo instrumento de dominio

A la luz de la crítica de Horkheimer, no cabe duda de cuál es el interés que tienen por la IA fuerte los grupos de poder que pueden costear su desarrollo: el dominio. En su vertiente realista, es decir, la IA humana, serviría para comprender el funcionamiento de la mente o del cerebro. Pero este enfoque, como hemos visto en el capítulo sexto, todavía está muy lejos de llegar a la simulación total de la cognición. Su ámbito es la IA débil, y no la fuerte, y por tanto no es cosa que nos ocupe aquí. El camino que más ha avanzado hacia la IA fuerte es el de la IA ajena, es decir, la vertiente instrumental, concebida para aplicaciones prácticas. Sobre tales aplicaciones Isaac Asimov observa lo siguiente: «Una computadora, no importa su costo, es de gran valor si puede hacer lo que los seres humanos no pueden hacer» (Asimov, 1996, p. 29). Y se pregunta a continuación: «¿Querría yo una computadora, diseñada por un costo enorme y siempre con el peligro de que se estropee, simplemente para escribir historias y ensayos para mí, cuando soy capaz de hacerlo yo mismo tan fácilmente (usando sólo lápiz y papel, si es necesario)?» (Ibíd., p. 29).

Especular sobre las aplicaciones concretas de la IA fuerte en manos de las corporaciones y los Estados es tarea que compete a la ciencia ficción, ya sea escrita a mano o a máquina, por lo que no vamos a hacer pronósticos al respecto. Basta con afirmar el enunciado general de que se utilizará para aumentar del dominio de los hombres sobre la naturaleza y sobre los propios hombres. Si se consiguen tales máquinas, sus dueños las emplearán para obtener de ellas lo que no puedan obtener

explotando a otros hombres. Así como la máquina de vapor se utilizó para realizar trabajos físicos imposibles para las bestias de carga, la IA fuerte se utilizará para realizar trabajos intelectuales inasequibles o no rentables si fueran realizados por seres humanos. Un ejemplo es la máquina de escuchas telefónicas temida por Weizenbaum, a la que venimos refiriéndonos desde el primer capítulo (Weizenbaum, 1976, p. 223). El temor de este filósofo e ingeniero era propio de los años 70, cuando las telecomunicaciones ordinarias eran por teléfono y por fax. En la actualidad la máquina de Weizenbaum serviría no sólo para pinchar teléfonos, sino para filtrar los más de 300.000 millones de correos electrónicos que se envían cada día en el mundo. Lo más parecido que existe a esta IA fuerte es el programa PRISM. Por culpa de un espía con mala conciencia, el presidente de los Estados Unidos se vio obligado a admitir públicamente en 2013 que varias agencias de inteligencia del país utilizaban PRISM, contando para ello con la colaboración de grandes corporaciones informáticas.

La previsión más verosímil es que en una sociedad como la actual, dirigida por la razón subjetiva y huérfana de razón objetiva, la IA fuerte será utilizada por los grupos de poder que pueden costearla para aumentar su dominio sobre el resto de la población. Ahora bien, el incremento del dominio *puede* tener un límite. Subrayemos que puede, pero no es necesario que lo tenga. Veamos por qué.

El cambio en la composición de valor

Mientras comenta el análisis del proceso del trabajo de Marx, Marcuse señala lo siguiente: «Los avances de la técnica disminuyen la cantidad de trabajo vivo (factor subjetivo) empleado en el proceso productivo, en proporción con la cantidad de los medios de producción (factor objetivo). El factor objetivo aumenta al disminuir el factor subjetivo. Este cambio en la composición técnica del capital se refleja en el cambio de su "composición de valor": el valor de la fuerza de trabajo disminuye al aumentar el valor de los medios de producción» (Marcuse, 1941, p. 303). Esto quiere decir que la automatización de la producción da lugar a una reducción del valor del

trabajo del obrero. Si fabricar una bota a mano en la Edad Media requería diez horas de tiempo de trabajo, hacerlo a máquina después de la Revolución Industrial requiere la décima parte, por lo que el obrero, para seguir ganando lo mismo, tiene que producir diez veces más. Y lo mismo se puede decir de la automatización del trabajo intelectual gracias a técnicas como las redes de computadores humanos de De Prony y Maskeline, de las que hablamos en el capítulo segundo, y gracias a tecnologías como las computadoras electrónicas y la IA fuerte.

A cierta escala, este cambio en la composición de valor es beneficioso para el obrero, pues a nivel individual trabaja el mismo tiempo y gana lo mismo, mientras que los productos son más abundantes y baratos, por lo que puede consumirlos en mayor cantidad. Sin embargo tiene dos consecuencias negativas. La primera es que los pequeños productores terminan desapareciendo, debido a que no pueden competir con los precios de los grandes, mucho más reducidos gracias a su posesión de la técnica más moderna para la automatización de la producción. En consecuencia: «La competencia individual libre, de estampa liberal, se transforma en una competencia monopolista entre empresas gigantes» (Ibíd., p. 303). Ésta es la explicación económica de por qué, como decíamos al comienzo del presente capítulo, la figura del pequeño empresario independiente ya no es típica (Horkheimer, 1947, p. 152), y su lugar ha sido ocupado por las grandes corporaciones. La segunda consecuencia es que, al necesitarse menos cantidad de trabajo vivo para producir lo mismo, son muchos los obreros que se quedan sin empleo. El único remedio para mantener a la masa obrera ocupada es producir más y consumir más. Así es como la economía capitalista entra en una espiral infinita de aumento de la producción.

Dado que los recursos del planeta en el que vivimos son finitos, la creciente automatización de la producción pondrá, tarde o temprano, al capitalismo ante una disyuntiva. La primera opción es la de mantener la duración de la jornada laboral y reducir progresivamente la cantidad de obreros empleados. La segunda es la opuesta: reducir la duración de la jornada laboral y mantener la cantidad de obreros empleados. Acerca de la primera, daría lugar a un aumento de la población que vive en la miseria.

Aumentaría, por tanto, la fuerza del sujeto que ha de realizar en algún momento la revolución hacia la pacificación de la existencia (Marcuse, 1964, p. 41), es decir, hacia una sociedad no represiva. Como decía Walter Benjamin: «Sólo gracias a aquellos sin esperanza nos es dada la esperanza» (Ibíd., p. 222). El protagonista de la novela de Orwell *1984* decía lo mismo que Benjamin, pero con otras palabras: «*Si hay alguna esperanza*, escribió Winston, *está en los proles*. Si había esperanza, tenía que estar en los proles porque sólo en aquellas masa abandonadas, que constituían el ochenta y cinco por ciento de la población de Oceanía, podría encontrarse la fuerza suficiente para destruir al Partido» (Orwell, 1949, p. 81). Ahora bien, el aumento de la miseria sólo aportaría la condición material de la revolución. Para que ésta se desatase faltaría además que el sujeto revolucionario cumpliese con la condición intelectual de tener *conciencia* de la *posibilidad material* y de la *necesidad moral* del cambio. De este asunto nos ocuparemos tras examinar la otra opción del capitalismo ante el crecimiento de la automatización de la producción.

Esta segunda opción sería, como decimos, la de reducir la duración de la jornada laboral y mantener la cantidad de obreros empleados, con salarios que les permitieran mantener un nivel de vida similar al actual. Sobre esta segunda alternativa Marcuse apunta lo siguiente: «Dentro de las sociedades establecidas, la aplicación continuada de la racionalidad científica alcanzará un punto final con la mecanización de todo el trabajo socialmente necesario pero individualmente represivo (el término "socialmente necesario" incluye aquí todas las acciones que pueden ejercerse con mayor efectividad por máquinas, incluso si estas actuaciones producen lujos y despilfarro más que necesidades). Pero este estadio será también el fin y el límite de la racionalidad científica en su estructura y dirección establecidas. El progreso ulterior implicaría la *ruptura*, la conversión de la cantidad en calidad. Abriría la posibilidad de una realidad humana esencialmente nueva; la de la existencia en un tiempo libre sobre la base de las necesidades vitales satisfechas» (Marcuse, 1964, p. 202). Como vemos, esta segunda alternativa, aunque por otra vía, conduce a la misma situación que la anterior: a las condiciones materiales necesarias para transformar la estructura social

represiva bajo la que vivimos en la actualidad por otra no represiva en la que el tiempo de trabajo fuese marginal y, en cambio, abundase el tiempo libre. Sin embargo, vuelve a suceder lo mismo que antes, y es que las condiciones materiales por sí solas no son suficientes para generar un cambio cualitativo de semejante envergadura. Haría falta, además, que se diesen unas ciertas condiciones intelectuales para que los individuos de tal hipotético escenario empleasen el tiempo libre en autorrealizarse como seres humanos, en lugar de consumirlo en forma de tiempo de ocio. Porque el tiempo libre y el tiempo de ocio no son la misma cosa (Ibíd., p. 219).

Así pues, la creciente automatización de la producción, a la cual la IA fuerte contribuiría en el ámbito del trabajo intelectual no menos de lo que lo hizo en su día la máquina de vapor en el físico, conduce a dos escenarios posibles que coinciden en poner las bases materiales para una revolución social, pero que, por sí solos, carecen de las condiciones intelectuales requeridas. Por eso hemos subrayado antes que el incremento del dominio puede tener un límite, pero no es necesario que lo tenga. Depende de que se den o no las citadas condiciones intelectuales. Si se dieran, el dominio caería. De lo contrario, seguiría vigente y en aumento, aún cuando se diesen las condiciones materiales para su aniquilación.

La cultura de masas

El principal obstáculo que impide que se den las condiciones intelectuales necesarias para trascender la sociedad actual hacia una sociedad no represiva es la *cultura de masas*, razón por la cual es uno de los temas centrales de la escuela de Frankfurt (Horkheimer, 1947, p. 43). Pensemos en aquellos sin esperanza a los que se refiere Benjamin. Por desgracia, lo que pretenden no es destruir las fronteras que separan a los países ricos de los pobres, sino atravesarlas. Desean con fervor participar en la sociedad de consumo que han visto en la televisión. Los techos de las favelas de están densamente cubiertos de antenas parabólicas para captar la señal de la cultura de masas de los países ricos: películas, seriales y competiciones deportivas.

Si los más perjudicados por el sistema, que son los de fuera, no dan signos de esperanza, la situación de los de dentro es todavía más desalentadora, porque ni siquiera pasan hambre. Desarrollan sus vidas inmersos en un aparato productivo que, como dice Marcuse, «tiende a hacerse totalitario en el grado en que determina, no sólo las ocupaciones, aptitudes y actitudes socialmente necesarias, sino también las necesidades y aspiraciones individuales» (Marcuse, 1964, p. 24). Cuando escribió estas palabras, en los años 60, Marcuse contemplaba los primeros pasos del proceso de extensión de este totalitarismo hacia el tercer mundo. Tal proyecto de dominación es hoy una realidad consolidada. El inmigrante llega al país rico sabiendo ya la marca de coche que quiere, el equipo de fútbol del que va a hacerse socio y el modelo de teléfono móvil que le gustaría tener. Gracias al alcance global de la cultura de masas, sus deseos y aspiraciones son los mismos de aquellos que ya están dentro.

Steve Jobs, fundador de la empresa informática Apple, es uno de los personajes que más han contribuido a modelar al individuo actual en todas partes del mundo: desde su país, los Estados Unidos, hasta Turquía. En un viaje a Estambul, Jobs contrató a un profesor de Historia. Mientras éste le explicaba las particularidades del café turco, Jobs tuvo una revelación: «¿A qué chicos, incluso en Turquía, les importa una mierda el café turco? Llevaba todo el día viendo jóvenes en Estambul. Todos bebían lo que beben los demás chicos del mundo, todos llevaban ropa que parecía sacada de una tienda Gap y todos utilizaban teléfonos móviles. Eran iguales que los jóvenes de todas partes. Me di cuenta de que, para los jóvenes, el mundo entero es un mismo lugar. Cuando fabricamos nuestros productos no pensamos en un "teléfono turco", o en un reproductor de música que los jóvenes turcos quieran y que sea diferente del que cualquier joven del resto del mundo pueda querer. Ahora somos todos un mismo planeta» (Isaacson, 2011, p. 724). El individuo del mundo globalizado desea los mismos productos. Pueden ser de diferentes marcas, o incluso con diferentes especificaciones técnicas, pero en el fondo son lo mismo. «Incluso entre los tipos más caros y los más baratos de la colección de una misma firma, las diferencias tienden a reducirse cada vez más» (Horkheimer & Adorno, 1944, p. 168).

En cierto modo, se podría decir que la cultura de masas *introyecta* en el individuo lo que éste debe pensar y desear. Sin embargo, Marcuse va más allá en su crítica, y rechaza la metáfora de la introyección en tanto que ésta sugiere la existencia de dos dimensiones distintas: una exterior desde la que se introyecta, y otra interior que recibe el contenido. En realidad, dice, ya no existe una dimensión interior distinta a la exterior (Marcuse, 1964, p. 36). De ahí el título de su obra más célebre, *El hombre unidimensional*. El sujeto se ha mimetizado con la sociedad. En él no hay nada que no le venga impuesto por ella. La *mímesis*, señala Horkheimer, es el más antiguo mecanismo biológico de supervivencia (Horkheimer, 1947, p. 153). Ciertamente, en todas las épocas ha existido la necesidad de adaptación. La diferencia distintiva de la actual es que, más que una adaptación, la cultura de masas opera una asimilación total (Ibíd., p. 121). En el pasado, la conciencia de los trabajadores era infraevolucionada, pero al menos «no estaba expuesta al constante acoso de las técnicas de la cultura de masas, que inculcan a sangre y fuego los patrones industrializados de conducta a su ojos, a sus oídos y a sus músculos, tanto durante su tiempo libre como durante su jornada laboral» (Ibíd., p. 159). El objetivo de la cultura de masas es «cerrar los sentidos de los hombres, desde la salida de la fábrica por la tarde hasta la llegada, al día siguiente, al reloj de control» (Horkheimer & Adorno, 1944, p. 176).

Las características que los hombres sometidos a este sistema cultural totalitario adquieren de la sociedad con la que se mimetizan son, obviamente, los mismos que hemos dicho que caracterizan a ésta: la conducta psicopática y la esquizofrenia. Los individuos así forjados carecen de la dimensión interior donde se desarrolla el pensamiento negativo sobre el que se construye la identidad (Marcuse, 1964, p. 36). El *sí mismo*, observan Horkheimer y Adorno, es el resultado de la resistencia a lo otro (Horkheimer & Adorno, 1944, p. 107). Ahora bien, para poder resistirse es condición necesaria que haya una dimensión interior que oponga resistencia a lo exterior, es decir, una diferencia entre el sujeto y el objeto. La mímesis, en tanto que destructora de la dimensión interior, identifica a sujeto y objeto, eliminando tal diferencia. Así se cumple el viejo deseo de eliminar la tensión entre ambos polos.

El abismo kantiano finalmente se ha cerrado (Marcuse, 1941, p. 67). El sujeto, por fin, se reconoce en el objeto, como era la intención de Hegel. Pero no de la forma que él pretendía, y que pretende también cualquier filosofía auténtica, a saber, la de reconocerse el sujeto en el objeto porque éste ha sido modelado según la verdad objetiva descubierta racionalmente por aquél (Ibíd., p. 13), sino todo lo contrario: porque el sujeto ha renunciado a la resistencia. Así, se reconoce en el mundo por la sencilla razón de que él no es más que una instanciación, un reflejo de ese mundo. El sujeto se ha vuelto fungible. No es más que un ejemplar: «La industria cultural ha realizado malignamente al hombre como ser genérico» (Horkheimer & Adorno, 1944, p. 190). El hombre, guiado por la sola razón subjetiva, ha creado un sistema que ha hecho con él mismo lo que él había hecho previamente a la naturaleza: cosificarlo.

La astucia de Odiseo

A pesar de converger en este diagnóstico terrible, Horkheimer y Marcuse albergan esperanzas en ilustrar a la Ilustración y superar la barbarie. Para lograrlo, es necesario sacar a los esclavos de la caverna de Platón, una tarea difícil, pues habrá de ser realizada en contra de la voluntad de los propios sujetos a liberar, que han sido educados para ser felices en su condición de esclavos (Marcuse, 1964, p. 59). Sigue habiendo clases, pero la lucha de clases ha terminado. El reparto de bienes a una escala cada vez mayor (Ibíd., p. 23) y el poder mimético de la cultura de masas han conseguido unir a los antiguos antagonistas, que comparten ahora «un interés absoluto en la preservación y el mejoramiento del *statu quo* institucional» (Ibíd., p. 22). El objetivo de la teoría crítica es reactivar la lucha, algo que sólo conseguirá convenciendo a los explotados de la verdad que hay en ella. Esa verdad puede resumirse en un único juicio: «La forma básica de la economía de mercancías históricamente dada, sobre la cual reposa la historia moderna, encierra en sí misma los antagonismos internos y externos de la época, los renueva constantemente de una manera agudizada, y que, tras un período de ascenso, de desarrollo de fuerzas

humanas, de emancipación del individuo, tras una fabulosa expansión del poder del hombre sobre la naturaleza, termina impidiendo la continuación de ese desarrollo y lleva a la humanidad hacia una nueva barbarie» (Horkheimer, 1941, p. 257). Horkheimer puntualiza que es precisamente porque los explotados se oponen al cambio cualitativo de la sociedad por lo que se requiere una teoría, pues, de lo contrario, «ella sería algo espontáneo en sus beneficiarios» (Ibíd., p. 252).

Tal tarea de concienciación de la posibilidad y la necesidad del cambio corresponde a una *intelligentsia*, una clase social preparada a tal efecto. ¿Cómo ha de hacerlo? Ésta es la gran pregunta. Horkheimer se opone a la propaganda, pues dice que: «La filosofía no puede ser convertida en propaganda, ni siquiera de cara a los más nobles fines» (Horkheimer, 1947, p. 185). Otra manera, paradójica en tanto que es la propia de la razón subjetiva que nos ha llevado a la situación actual, sería utilizando la *astucia de Odiseo*. Ésta consiste en satisfacer «la norma jurídica de tal forma que ésta pierde poder sobre él (el sujeto que la satisface) en el momento mismo en que él se lo reconoce» (Horkheimer & Adorno, 1944, p. 110). Odiseo ordena a su tripulación que lo amarren al mástil, y que ellos se tapen los oídos con cera. De esa manera, el héroe cumple con la norma impuesta por las sirenas, que es escuchar su canto, pero burla el espíritu de la norma, que es conseguir que los marineros se suiciden.

La norma que impone la industria cultural es la rentabilidad de la mercancía, para lo cual ésta debe resultar atractiva a los gustos condicionados del consumidor. La astucia aplicada a la industria cultural consistiría, por tanto, en producir obras que satisficieran la norma, procurando grandes cifras de ventas, al tiempo que, por su contenido, atentasen contra el sistema sustentado por ella. Un ejemplo de astucia de este tipo sería *1984*. El cineasta Michael Moore habla sobre la astucia de Odiseo parafraseando a Lenin en los últimos minutos del documental *The corporation*. La cita es un poco larga, pero merece la pena, ya que se trata del testimonio de uno de los más exitosos divulgadores del pensamiento crítico: «¿Saben? Siempre he pensado en lo irónico que es que pueda hacer todo esto y seguir dedicándome a lo mío. Trabajo en la industria audiovisual. Los estudios me distribuyen y son propiedad de grandes

corporaciones. Entonces, ¿por qué me siguen comprando cuando resulta que me opongo a todo aquello que ellos representan? ¿Cómo es que empleo el dinero que me dan ellos para oponerme a aquello en lo que ellos creen? Pues porque no creen en nada. Me distribuyen porque saben que hay millones de personas que quieren ver mi película o ver el programa de televisión, y ellos van a ganar dinero. Y yo he conseguido distribuir mi material gracias a que voy conduciendo mi camión por esta enorme grieta que existe en el capitalismo: la grieta de la *codicia*. Es como el hombre rico que te vende la cuerda para ahorcarle si cree que con ello va a hacer dinero. Pues bien; yo soy esa cuerda. Soy parte de esa cuerda. También creen que cuando la gente ve mi trabajo o vean este documental, o lo que sea, se creen que ustedes lo verán y no harán nada al respecto porque ellos han conseguido ya anular sus mentes y ustedes no podrán reaccionar; no se van a levantar de sus sillones para cambiar las cosas. Están convencidos de ello. Y yo estoy convencido de lo contrario: estoy convencido de que algunas personas van a salir del cine o levantarse del sillón y van a hacer algo, cualquier cosa, para recuperar el mundo y volver a tenerlo en nuestras manos».

Horkheimer y Adorno, sin embargo, están convencidos de que semejante esfuerzo de divulgación mediante la astucia de Odiseo está condenado al fracaso. Su argumento es que el consumidor de la cultura de masas está condicionado hasta tal punto que es incapaz de entender nada que vaya más allá del discurso establecido. En términos kantianos, lo expresan así: «La tarea que el esquematismo kantiano esperaba aún de los sujetos, a saber, la de referir por anticipado la multiplicidad sensible a los conceptos fundamentales, le es quitada al sujeto por la industria (cultural). Para el consumidor no hay nada por clasificar que no haya sido ya anticipado en el esquematismo de la producción» (Horkheimer & Adorno, 1944, p. 169). Y lo rematan con dos sentencias lapidarias. Ésta es una: «El mundo entero es conducido a través del filtro de la industria cultural» (Ibíd., p. 171). Y ésta es la otra: «Sentido de la realidad, adaptación al poder, no son ya resultado de un proceso dialéctico entre el sujeto y la realidad, sino producidos directamente por el mecanismo industrial. [...] Al individuo [...] lo han extinguido como sujeto» (Ibíd., p. 248).

Por desgracia, la evidencia nos obliga a estar de acuerdo con Horkheimer y Adorno. Sobre el papel, la astucia de Odiseo es eficaz, pero en la práctica no. A raíz de la revelación del programa PRISM antes mencionado, *1984* se convirtió durante unos días en el libro más vendido en los Estados Unidos. Sin embargo, sería ingenuo creer que en los próximos comicios electorales la novela de Orwell tendrá algún efecto notable sobre el voto de quienes la leyeron. Lo más probable es que la lean y no la entiendan. Y si la entienden, la *doblepensarán*, igual que *doblepiensan* el sermón del cura o las campañas de recogida de alimentos para el tercer mundo. Han sido educados desde la infancia en esa conducta esquizofrénica. Trescientas cincuenta páginas leídas en la cama, después de una jornada laboral que ha consumido todas las energías, no tienen la fuerza suficiente para competir con la televisión, la radio, las revistas, los periódicos, y sus toneladas de propaganda adoctrinadora.

La industria cultural ha tapado los oídos de los consumidores. Éstos son como los remeros de Odiseo. De nada sirve hablarles, porque no pueden escuchar. En su aislamiento monádico sólo se preocupan de remar con fuerza para que la nave siga avanzando hacia el rumbo ordenado por el patrón. Es una metáfora muy del gusto de los políticos. Para salvar a la sociedad de la miseria en que ellos mismos la han hundido, ya sea por simple incompetencia o por orden de las corporaciones para las que verdaderamente trabajan, dicen que es necesario alcanzar grandes pactos de Estado. Hay que aunar esfuerzos, proclaman, aparcar las diferencias y remar todos en la misma dirección. La industria cultural, con la fuerza de sus medios de comunicación, se encarga de que así sea. Es el capataz que marca el ritmo en la galera.

Una prueba de la confianza que tienen los grupos de poder en que esta situación es irreversible la encontramos en *El hombre unidimensional*. Pero no en el texto general, sino en una sección que suele pasarse por alto: la de agradecimientos. Allí Marcuse agradece, entre otras, a la Rockefeller Foundation que le otorgase becas para escribir ese libro (Marcuse, 1964, p. 17). La Rockefeller Foundation fue creada en 1913 por John Davidson Rockefeller, el mismo magnate del petróleo que dos décadas antes había fundado la universidad de Chicago, conocida por ser la casa de una escuela

económica liberal. El mismo capital, por tanto, tan pronto pone la primera piedra de una escuela de liberales como financia las investigaciones de un marxista que, por aquel entonces, los 60, tenía gran ascendencia sobre buena parte de los universitarios norteamericanos. Este hecho muestra hasta qué punto los grupos de poder están tranquilos produciendo las películas de Michael Moore, becando a Marcuse o imprimiendo nuevas ediciones de *1984*. Están tranquilos porque están convencidos que no están vendiendo la soga con la que los van a colgar. Y están convencidos de ello porque saben que, gracias a la industria cultural, el poder transgresor de todas esas obras será inmediatamente anulado, reduciéndolas a simples mercancías. A la naturaleza se la vence obedeciéndola, decía Bacon. Pero a la industria cultural no, porque ella es inmune a la astucia de Odiseo.

En conclusión, la IA fuerte no dará lugar a ningún cambio cualitativo por sí sola. El cambio en la composición de valor, que será agudizado por ella si algún día se logra, no será suficiente para trascender la sociedad actual, como tampoco fueron suficientes en el pasado la máquina de vapor y la computadora electrónica. Como dice Javier Bustamante citando a Salvador Giner, las computadoras fueron una *revolución sin revolución* (Bustamante, 1993, p. 39). Lo mismo pronosticamos nosotros sobre la IA fuerte en caso de que su posibilidad técnica se haga efectiva. La razón es que, además de condiciones materiales, deben darse condiciones intelectuales, las cuales son imposibles debido al adoctrinamiento totalitario que la sociedad industrial avanzada ejerce sobre el individuo, especialmente a través de la industria cultural.

9. Conclusión

El análisis de las condiciones de posibilidad técnicas de la IA fuerte nos ha conducido a la conclusión de la imposibilidad por principio de la IA simbólica. La IA se basa en cuatro supuestos que, en la terminología de Lakatos, se ordenan como sigue desde el cinturón hacia el núcleo: biológico, psicológico, epistemológico y ontológico. Todos ellos son falsos. El supuesto biológico, según el cual el cerebro es una máquina de estado discreto equivalente a una computadora electrónica, encontró un amplio apoyo en los años 50 gracias a la similitud entre el carácter binario de la información computacional y el carácter también binario, o más correctamente booleano, de todo o nada, de los potenciales de acción con los que se comunican las neuronas. Sin embargo, hemos visto cómo el propio John von Neumann, uno de los padres de la informática, puso en duda en aquella época el supuesto biológico al señalar los factores analógicos cruciales de la comunicación entre las neuronas. Del contraste entre la descripción de las propiedades formales de las computadoras electrónicas realizada en el capítulo tercero y los conceptos fundamentales de neurociencia expuestos en el quinto ha resultado evidente que el cerebro no es una máquina digital, sino analógica, es decir, que la información con la que opera está representada no por elementos discretos, sino por variables físicas continuas.

El supuesto psicológico enuncia lo mismo que la hipótesis fuerte del sistema de símbolos (HFSS), a saber: que la mente utiliza procesos computacionales para producir la conducta inteligente. De esta manera, la posibilidad técnica de la IA simbólica estaría garantizada por principio, ya que la mente humana no sería más que un tipo de programa informático. Según Dreyfus el supuesto psicológico se fundamenta, a grandes rasgos, en una doble confusión: confundir el término "información" en su

sentido computacional y vulgar, y confundir la computación en general con la computación dirigida por reglas. Respecto a la primer confusión, la teoría de la información de Shannon y Weaver, que es el principio de la informática, establece que toda información, con independencia de su contenido, es codificable en alternativas binarias, es decir, en bits. Sin embargo, esta definición concibe la información en términos sintácticos, mientras que la noción vulgar de información implica contenidos semánticos, y la sintaxis, por definición, no es constitutiva ni suficiente para la semántica, pues la sintaxis se ocupa de la forma de los signos, y la semántica, de su contenido. En cuanto a la segunda confusión, es la que identifica la computación en general con la computación dirigida por reglas. En el cerebro, ciertamente, se realizan tareas de computación a nivel neuronal descriptibles por ciertas reglas descubiertas por la neurociencia, pero en la escala global del cerebro no hay reglas que describan la computación. Y, aunque las hubiera, eso no garantizaría que también las hubiese en el nivel de la mente, pues mente y cerebro no son lo mismo. Es verdad que algunos procesos mentales, que no todos, son descriptibles por reglas, pero de ahí no se puede concluir que la mente los produzca aplicando reglas. De la misma forma, el movimiento de los planetas, por ejemplo, es descriptible por reglas, pero de ello no se puede deducir legítimamente que los planetas se muevan resolviendo ecuaciones. El supuesto psicológico confunde el ser descriptible por reglas y el ser gobernado por ellas, como si lo primero implicara lo segundo.

El supuesto epistemológico sostiene lo mismo que la hipótesis del sistema de símbolos (HSS), a saber: que un sistema de símbolos físico tiene las capacidades suficientes para la acción inteligente general. Los padres fundadores de la IA, Allen Newell y Herbert Simon, proclamaron esta hipótesis como la ley general de la IA, de la misma manera que la doctrina celular es la ley general de la biología y la tectónica de placas lo es de la geología (Newell & Simon, 1975, p. 38). Sin embargo, en la HSS hay dos tesis implícitas que pueden ser refutadas: que toda conducta inteligente puede ser formalizada y que dicha formalización es, no necesaria como reivindica el supuesto psicológico, pero sí suficiente para reproducir la conducta en cuestión. La creencia

optimista de que toda conducta inteligente puede ser formalizada hunde sus raíces en la confianza positivista de que no hay nada que no pueda ser explicado de manera nomológico-deductiva, es decir, al estilo de la física. Este supuesto poder ilimitado debería por tanto extenderse también a la pragmática, que es la parte de la semiótica que estudia cómo el contexto afecta al significado del lenguaje. Sin conocimiento de la pragmática, no hay comprensión posible del lenguaje natural. Contra la posibilidad de formalizar la pragmática, y por tanto contra la posibilidad de que una computadora comprenda el lenguaje natural, Dreyfus expone dos argumentos distintos. Uno es un argumento de principio que se reserva para la refutación del supuesto ontológico. El otro es una descripción objetiva, que apela a la evidencia, sostenida también por el segundo Wittgenstein, de que no todas las conductas lingüísticas están regidas por reglas, algo que hemos ilustrado con dos textos de Julio Cortázar que violan respectivamente las reglas de la sintaxis y de la semántica pero, no obstante, son significativos. Si la comprensión del lenguaje se basara en la aplicación de reglas, entonces el sujeto debería disponer de reglas que explicasen las violaciones de las reglas, y así sucesivamente hasta el infinito.

En cuanto al supuesto ontológico, constituye el núcleo más duro del programa de investigación de la IA simbólica, tanto humana como ajena. Lo que sostiene es lo mismo que el atomismo lógico de Bertrand Russell y del primer Wittgenstein, a saber: que el mundo entero es expresable en una gran masa de hechos discretos y compuestos en última instancia de hechos simples autoevidentes. O dicho de otra manera: que el mundo entero es descriptible mediante sistemas formales. No sólo el mundo físico, que en efecto sí parece serlo desde un punto de vista instrumental a tenor del éxito de la física y otras ciencias de la naturaleza, sino el mundo social, el del espíritu. Dando por verdadero este supuesto, al ser las computadoras electrónicas sistemas formales, podrían por tanto comprender el mundo. Sin embargo, frente a esta epistemología objetivista del atomismo lógico, que postula una ingenua relación unidireccional de la información desde el objeto hasta el sujeto, nosotros hemos argumentado en favor de una epistemología relativista, bidireccional y holista basada

en el pragmatismo del segundo Wittgenstein y en la estructura circular de la comprensión descrita por Heidegger. Todo comprender, incluso el de los hechos supuestamente más simples, se funda en un conocimiento previo. Ahora bien, la selección del conocimiento previo pertinente depende de cuál sea el fenómeno a comprender, pero la identidad de éste depende a su vez del conocimiento previo desde el cual sea comprendido. Esta relación circular entre el todo y las partes resulta inasequible para las computadoras electrónicas, razón por la cual la IA simbólica, tanto humana como ajena, es imposible en definitiva.

Lo que pretende la IA simbólica, así como el cognitivismo en tanto que es su contrapartida en la psicología, es formalizar justamente aquello en virtud de cuya ignorancia obtienen las ciencias formales su potencia predictiva: el mundo social. Las fórmulas con que operan las ciencias de la naturaleza renuncian al sentido que hay en la dimensión social de la realidad para obtener a cambio el poder de la manipulación de la dimensión física. Explicar, como pretenden la IA simbólica y el cognitivismo, la dimensión social o mundo en los términos nomológico-deductivos en los que se explica la dimensión física o universo es la última tarea del positivismo; una tarea imposible. Cierto es que en el mundo, así como en el universo, hay regularidades, pues de lo contrario no podríamos habérnoslas en ellos. Sin embargo, sus reglas tienen una cantidad indefinida de condiciones de validez. Esto es algo que desde dentro de la IA supieron ver McCarthy y Hayes en 1969, pero que muchos investigadores, como el director del CYC Douglas Lenat, despreciaron y despreciarán siempre, pues es un argumento definitivo que refuta su intento alquímico de convertir el plomo en oro.

La forma en que los seres humanos, y en general todos los animales, nos las habemos con las regularidades tanto del mundo como del universo es mediante las memorias procedimentales. En ellas no hay nada simbólico, es decir, ninguna manipulación de representaciones cuasi-lingüísticas, sino sólo patrones de activación neuronal que representan el conocimiento de manera implícita y distribuida. Estos patrones sí son codificables en redes de neuronas artificiales, razón por la que hemos concluido que la IA subsimbólica es técnicamente posible. Cuál sea el nivel de detalle al

que deba ser replicado el funcionamiento de las redes de neuronas es una cuestión empírica sobre la cual no podemos pronunciarnos desde la filosofía. Lo que sí podemos hacer, y hemos hecho desde este saber, es señalar la necesidad de que la duplicación artificial del cerebro vaya acompañada de una duplicación del cuerpo. El cuerpo es la puerta de entrada y salida de la información neuronal, por lo que su participación en el modelado del cerebro es tan imprescindible como la del medio ambiente.

Junto con el cuerpo, para la creación de la IA fuerte es imprescindible también la duplicación del sistema motivacional. De ello hemos dado cuenta describiendo en el capítulo cuarto el funcionamiento del sistema nervioso central, compuesto de tres grandes sistemas funcionales estrechamente vinculados entre sí: sensorial, motor y motivacional. La motivación afecta a la percepción así como a la calidad de la ejecución incluso de las salidas motoras más simples (Kandel, Schwartz & Jessell, 1995, p. 84). En la IA simbólica la eliminación del estudio de las emociones responde, como dice Gardner, a la necesidad metodológica de eludir en sus inicios históricos la explicación de fenómenos mentales complejos, como son las emociones, postergando su inclusión a fases más avanzadas de la heurística positiva, las cuales, por otra parte, nunca han llegado ni llegarán jamás por ser el cognitivismo un paradigma basado en la ontología falsa, hace tiempo superada por la filosofía, del atomismo lógico. En cuanto a la IA simbólica, en ella también encontramos a investigadores que, como Jeff Hawkins, excluyen de sus intentos de duplicación artificial del pensamiento al sistema límbico en favor de una concepción netamente corticalista de la inteligencia.

Que Platón y Descartes, por citar a dos exponentes del racionalismo, separasen la razón y las pasiones obedecía a una cuestión fundamental para la filosofía que no debería traspasar a la ciencia pero que, sin embargo, lo ha hecho y sigue haciéndolo en el siglo XXI. En la filosofía, la separación de la razón y las pasiones es la respuesta a la necesidad de identificar respectivamente dos fuerzas contrarias: la de la construcción de la subjetividad y la de la disolución que amenaza con devolver al sujeto al estado de naturaleza salvaje. Horkheimer y Adorno lo explican a través de la *Odisea* de Homero. La odisea desde Troya a Ítaca, dicen, es el itinerario del sí mismo que todo hombre y

sociedad recorre a lo largo de su vida (Horkheimer & Adorno, 1944, p. 100). El sujeto encuentra en el camino numerosas tentaciones que lo invitan a abandonarse al placer y renunciar al viaje, que es largo y fatigoso. La respuesta de la razón es reprimir el instinto, renunciar a la satisfacción presente en aras de un bien mayor futuro. Hambre para hoy y pan para mañana: ésta es la máxima por la que se rige la construcción del sujeto. La felicidad no es inmediata, sino que es esencialmente un resultado: «Se desarrolla en y desde el dolor superado. Por ello, el paciente héroe está en su derecho, que no le permite quedarse entre los lotófagos» (Ibíd., p. 114). Los lotófagos son hombres que se nutren del loto, flores estupefacientes que dan una sensación de felicidad que, en el fondo, es falsa porque en ella no hay lucha ni trabajo, sino sólo disfrute de un eterno presente que no conduce a un futuro mejor.

Cuando la ciencia hace suya la división filosófica entre razón y pasiones y en consecuencia cree posible construir máquinas pensantes puramente racionales, lo que está haciendo es cometer un grave error categorial que engendra modelos defectuosos de la mente, como el de Phineas Gage, de pura razón desconectada de las pasiones (Damasio, 1994, p. 166). Por otra parte, se da la interesante circunstancia de que este tipo de mente, el de la matriz de Gage, se caracteriza por la incapacidad para dirigir racionalmente la vida. Ya dijimos que Gage parecía perfectamente racional cuando era diagnosticado en un breve espacio de tiempo, pero que, en cambio, a largo plazo era incapaz de planificar. Tras el accidente que destruyó la conexión de su racionalidad cortical con la emocionalidad de la amígdala, pasó sus días deambulando de un lugar a otro, ganándose la vida de cualquier manera sin pensar en el futuro.

Salta a la vista la similaridad entre la forma en que Gage, los lotófagos y los psicópatas viven sus vidas. Todos ellos son incapaces de marcarse objetivos a largo plazo y sacrificar el presente para lograrlos. Las causas, no obstante, son inversas: el lotófago, porque se entrega a las pasiones sin el freno de la razón, mientras que Gage y los psicópatas son razón casi pura, con emociones muy empobrecidas. La matriz de Gage y la psicopatía comparten a nivel clínico una carencia de emociones. Así lo señala Damasio, quien describe como sigue a los psicópatas: «A veces son listos. El umbral al

que sus emociones afloran, cuando lo hacen, es tan alto que resultan inconmovibles y, a partir de sus propios informes, son insensibles e impasibles» (Ibíd., p. 210). Y a continuación los compara con la matriz de Gage: «En realidad, son otro ejemplo de un estado patológico en el que una reducción de la racionalidad viene acompañada por una disminución o ausencia de sentimientos» (Ibíd., p. 210). En definitiva, desde un punto de vista científico la inteligencia necesita a las pasiones tanto como a la razón.

En cuanto a las conclusiones sobre el examen de las condiciones de posibilidad sociales de la IA, volvamos al comienzo del primer capítulo. Allí definíamos la noción vulgar de IA, que coincide con la IA fuerte, como una máquina con una inteligencia parecida a la de un ser humano que por su condición de máquina se espera que sirva a su dueño, y por su condición de inteligente se espera que lo haga con unas destrezas intelectuales parecidas a las de un ser humano. Se trata de una definición abstraída de las múltiples formas en que la IA se presenta en la cultura popular, particularmente en la ciencia ficción, un género que, cuando es valioso, contiene reflexiones filosóficas.

Así, por ejemplo, Asimov habla de los robots del futuro para denunciar los males del presente. La raíz de todos los males del presente viene de atrás, como observan Horkheimer y Adorno, y es la hegemonía de una razón instrumental liberada de enfrentarse a los imperativos éticos y políticos de la razón objetiva. En una sociedad así, los individuos libres e iguales sólo existen en un plano legal que forma parte del encubrimiento esquizofrénico de la realidad. En la práctica, en una sociedad semejante los hombres se ven unos a otros como dice Hare que lo hacen los psicópatas: el mundo para ellos se divide en depredadores y víctimas. Las tres leyes de la robótica de Asimov son las tres normas que al depredador le gustaría imponer. La primera es que la víctima no debe dañar al depredador o por su inacción dejar que éste sufra daño alguno. La segunda es que la víctima debe obedecer las órdenes que le son dadas por el depredador, excepto cuando estas órdenes se opongan a la primera ley. Y la tercera establece que la víctima debe proteger su propia existencia, pero sólo hasta donde esa protección no entre en conflicto con las dos leyes anteriores. Las tres leyes de la robótica de Asimov son las leyes de la sociedad en que vivimos.

Dado que las inteligencias artificiales de la ciencia ficción son reflejo de la sociedad actual, y que la noción vulgar de IA está determinada por la ciencia ficción en retroalimentación con la mercadotecnia y las divulgaciones de los investigadores de la IA, lo que el público desea al desear la IA fuerte es una perpetuación de la estructura social existente. La fuerza negativa que, como arte que son, contienen las obras de los grandes creadores de la ciencia ficción, es neutralizada por la industria cultural, que las emplea como instrumento de reafirmación de la realidad. Ya señalamos en el primer capítulo que la ciencia ficción presenta a la IA de manera distópica, esto es, advirtiendo de los peligros que de ella pueden derivar, aunque en realidad más que advertir del futuro, como decimos, advierte del presente. Sin embargo, el público las desea. Las desea porque desea convertirse en depredador, en amo rodeado de esclavos. Como este sueño se le antoja imposible en la vida real, la ciencia ficción le ofrece, al igual que el cine de psicópatas, una sublimación de su deseo.

No es nada nuevo. El rabino Low no crea al golem de Praga para incorporarlo a la comunidad judía como un miembro más de pleno derecho, sino para utilizarlo como esclavo que defienda al gueto (Salfellner, 2011, p. 46). Una vez más, se cumplen las tres leyes de la robótica. El golem no debe dañar a los judíos ni dejar que éstos sean dañados, debe obedecerlos y debe proteger su propia vida, por este orden de prioridad. Los científicos de la IA pueden albergar pretensiones más elevadas, como la de desvelar el conectoma del cerebro humano para aplicaciones sanitarias. No obstante, no son ellos los que deciden los usos de sus descubrimientos, sino la sociedad que los financia. Esta sociedad, la nuestra, está dirigida por grupos de poder carentes de razón objetiva que a lo único que aspiran es a la ampliación de su dominio sobre el resto de la humanidad. Una vez sometida la naturaleza mediante la ciencia, ellos aspiran a convertirse en la nueva naturaleza (Horkheimer & Adorno, 1944, p. 157), fuente de leyes inviolables que los hombres deben cumplir. Prueba del afán de dominación que ha guiado siempre a la IA es la importancia que la inversión militar ha tenido en su desarrollo. En 1974 se paralizó por un corte de la financiación de DARPA, la agencia estadounidense de investigación para la defensa del país.

El golem de Praga, recordemos, se volvió contra sus amos, causando la destrucción del gueto judío (Salfellner, 2011, p. 49). La técnica tiene un doble filo. Lo observamos también en la película *Dr. Strangelove* (1964) de Stanley Kubrick. En los años 60, en plena Guerra Fría, los Estados Unidos tienen una flotilla de bombarderos B-52 equipados con armas nucleares sobrevolando permanentemente las proximidades de Rusia para atacar en cuanto se les dé la orden. Un general del ejército fuera de sus cabales da, en efecto, la orden. Cuando la cúpula militar se reúne para revocarla, resulta que es imposible a causa del Plan R. El Plan R se diseñó precisamente para que nadie pudiera ponerse en contacto con los B-52 una vez iniciado el ataque. Ante la gravedad de la situación, el embajador ruso, allí presente con la cúpula del Pentágono, revela que su país tiene una respuesta preparada: la Máquina del Apocalipsis. Cuando esta máquina detecte el ataque nuclear, pondrá en marcha una contrarréplica también imparable que destruirá toda la vida sobre la faz de la Tierra. El presidente de los Estados Unidos se pregunta escandalizado cómo es posible semejante monstruosidad, a lo que el Dr. Strangelove le responde: «Señor presidente, no es únicamente posible; es *esencial*, es la idea en que se basa esta Máquina –aclara el doctor Strangelove, muy obviamente enardecido por poder comentar los aspectos teóricos de la cuestión–: la disuasión es el arte de producir en la mente del enemigo el *miedo* al ataque. Por lo tanto, como el proceso decisorio es automático e irrevocable y funciona fuera del control humano, la Máquina del Apocalipsis es terrible» (Rivera, 2003, p. 182).

La seguridad pretendida con el golem se torna en inseguridad. La paz pretendida con la Máquina del Apocalipsis se torna en guerra. La sabiduría pretendida con la escritura, advierte Platón en el *Fedro* como vimos en el capítulo tercero, se torna en ignorancia. La información pretendida con Internet, advierte Nicholas Carr como vimos en el capítulo quinto, se torna en el modelado de individuos carentes de pensamiento profundo, entendiendo por tal, dice Carr, aquel que comprende «la adquisición de conocimiento, el análisis inductivo, el pensamiento crítico, la imaginación y la reflexión» (Carr, 2010, p. 141). La técnica puede volverse contra aquel que la utiliza deparándole un resultado opuesto al que deseaba. Éste es el resquicio

que deja lugar a la esperanza. En el caso de la IA fuerte, la automatización de las tareas intelectuales tiene como objetivo el aumento de la productividad dentro del sistema económico actual. El resultado contrario que podría suceder es que, al condenar esa tecnología al desempleo a millones de personas, aumentase la población que vive en la miseria, por lo que aumentaría la fuerza del sujeto histórico que debe trascender el orden social imperante hacia otro más justo, guiado por la razón objetiva. Sin embargo, tal posibilidad, como hemos visto en el capítulo octavo, ha sido desactivada por la industria cultural. Ésta llega hasta el último rincón del planeta para ejercer un adoctrinamiento totalitario sobre los que no tienen esperanza, de tal forma que éstos, en lugar de aspirar a terminar con la sociedad de depredadores y víctimas, aspiran a convertirse en depredadores. El universo del discurso y de la acción ha sido cerrado de manera tan perfecta que ni siquiera la astucia de Odiseo, que es la esencia de la técnica que nos ha llevado a esta situación en tanto que es el principio de dominación de la naturaleza, puede darse la vuelta para emplear su otro filo.

Hemos visto que para Horkheimer, y nosotros coincidimos con él en este punto desde la teoría de las inteligencias múltiples de Howard Gardner, la inteligencia no es sólo razón subjetiva, sino también la capacidad para percibir contenidos objetivos, es decir, contenidos de la razón objetiva. La noción vulgar de IA no dice eso. Para el público la máquina inteligente es aquella que obedece ciegamente a su dueño. De lo contrario, si tuviera criterio propio para discernir el bien del mal y potestad para aplicarlo, no sería de gran utilidad. Lo que la sociedad quiere de la IA fuerte es lo mismo que quiere del obrero: sumisión. El obrero quiere creer que su lugar será ocupado por la máquina. Y, en efecto, lo será, pero él no obtendrá ningún beneficio de tal reemplazo, porque las máquinas sólo trabajan para sus propietarios. El sueño del obrero es profundamente positivista. Está convencido de que el solo avance de la técnica lo liberará, sin darse cuenta de que cuanto más avanza la técnica para liberarlo de la coerción de la naturaleza, más avanza el dominio de los grupos de poder sobre él, para ocupar, como decimos, el lugar dejado por la naturaleza. La técnica sólo opera la sustitución de un amo por otro. La servidumbre permanece.

En una sociedad como la actual, con un imparable aumento de la digitalización de la información, la única salvaguarda de la libertad del individuo que impide el advenimiento final del mundo de Orwell es la no-existencia de la IA fuerte. El enorme volumen de datos, en crecimiento exponencial, es lo que impide a los grupos de poder el control total sobre la vida. La IA fuerte eliminaría ese obstáculo. En virtud del aumento constante del tráfico de información digital no sólo en el sentido extensivo sino también, y sobre todo, en el intensivo de reflejar la totalidad de la vida, una computadora capaz de analizar todo ese tráfico sería cada vez más cercana a una deidad omnisciente. Pero una deidad sin voluntad propia, pues como dicta la noción vulgar de IA, por su condición de máquina se espera que sirva a su dueño. Aquellos que la controlasen, por tanto, controlarían el mundo con un grado de exhaustividad sin precedentes en la Historia. Si la Ilustración sigue su curso degenerado, es algo que tarde o temprano sucederá. Y no hay ningún indicio de que no vaya a ser así.

Bibliografía

Ananthanarayanan, R., Esser, S. K., Simon, H. D., & Modha, D. S. (2009). The cat is out of the bag: Cortical simulations with 10^9 neurons, 10^{13} synapses. *Proceedings of the ACM/IEEE conference on supercomputing* (Portland, OR, Nov. 14-20). ACM, New York, 1-12.

Anastasi, A. (1986). Intelligence as a quality of behavior. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 19-22).

Asimov, I. (1950). *Yo, robot*. Barcelona: Planeta; 2006.

Asimov, I. (1985). *Robots e Imperio*. Barcelona: Mondadori; 2007.

Asimov, I. (1996). *¿Cómo será el futuro?* Gerona: Tikal.

Baltes, P. B. (1986). Notes on the concept of intelligence. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 23-28).

Baron, J. (1986). Capacities, dispositions and rational thinking. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 29-34).

Berger, Th. W. (2005). Restoring lost cognition function. *IEEE Engineering in medicine and biology magazine, september/october*, 30-44.

Berry, J. W. (1986). A cross-cultural view of intelligence. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 35-38).

Black, E. (2001). *IBM and the holocaust: The strategic alliance between nazi Germany and America's most powerful corporation*. California: Three Rivers Press.

Brooks, R. A. (1991). Intelligence without representation. En J. Haugeland (Ed.), *Mind design II* (pp. 395-420).

Brown, A. L., & Campione, J. C. (1986). Academic intelligence and learning potential. En R. Sternberg & D. Detterman (Eds.), *What is intelligence?* (pp. 39-44).

Bunge, M. (1987). La psicología, ¿disciplina humanística, autónoma, natural o social? *Arbor, Ciencia, pensamiento y cultura*, 496, 9-30.

Bustamante, J. (1993). *Sociedad informatizada, ¿sociedad deshumanizada?* Madrid: Gaia.

Bustamante, E. (2007). *El sistema nervioso: Desde las neuronas hasta el cerebro humano*. Colombia: Editorial Universidad de Antioquía.

Butterfield, E. C. (1986). Intelligent action, learning and cognitive development might all be explained with the same theory. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 45-50).

Carpintero, H. (1996). *Historia de las ideas psicológicas*. Madrid: Pirámide; 1998.

Carr, N. G. (2010). *The shallows: How the Internet is changing the way we read, think and remember*. London: Atlantic Books.

Carroll, J. B. (1986). What is intelligence? En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 51-54).

Castellet, J. M. (1969). *Lectura de Marcuse*. Barcelona: Seix Barral.

Ceruzzi, P. E. (1998). *A history of modern computing*. Cambridge: The MIT Press: 2002.

Chalmers, A. F. (1982). *¿Qué es esa cosa llamada ciencia?* Madrid: Siglo XXI; 1991.

Curchland, P. M. (1989). On the nature of theories: A neurocomputational perspective. En J. Haugeland (Ed.), *Mind design II* (pp. 251-292).

Clark, A. (1992). The presence of a symbol. En J. Haugeland (Ed.), *Mind design II* (pp. 377-393).

Comte, A. *Curso de filosofía positiva: Lecciones I y II*. Buenos Aires: Ediciones Libertador; 2004.

Comte, A. *Discurso sobre el espíritu positivo*. Madrid: Alianza; 2000.

Copeland, B. J. (1993). *Inteligencia artificial: Una introducción filosófica*. Madrid: Alianza; 1996.

Copeland, B. J. (Ed.) (2004). *The essential Turing*. Oxford: Clarendon Press.

Coppin, B. (2004). *Artificial intelligence illuminated*. Massachusetts: Jones & Bartlett.

Crevier, D. (1993). *The tumultuous history of the research for AI*. New York: Basic Books.

Damasio, A. (1994). *El error de Descartes*. Barcelona: Crítica; 2007.

Damasio, A. (2000). *La sensación de lo que ocurre*. Madrid: Debate; 2001.

Damasio, A. (2003). *En busca de Spinoza*. Barcelona: Crítica; 2005.

Das, J. P. (1986). On definition of intelligence. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 55-56).

Davidson, D. (1973). The material mind. En J. Haugeland (Ed.), *Mind design* (pp. 339-354).

Davidson, J. E., & Kemp, I. A. (2011). Contemporary models of intelligence. En R. J. Sternberg (Ed.), *The Cambridge handbook of intelligence* (pp. 58-82).

Deitel, H. M., & Deitel, P. J. (2012). *C++: How to program (8th edition)*. Boston: Prentice Hall.

Dennett, D. C. (1971). Intentional systems. En J. Haugeland (Ed.), *Mind design* (pp. 220-242).

Dennett, D. C. (1981). True believers: The intentional strategy and why it works. En J. Haugeland (Ed.), *Mind design II* (pp. 57-79).

Descartes, R. *Discurso del método & Reglas para la dirección de la mente*. Barcelona: Orbis; 1983.

Detterman, D. K. (1986). Human intelligence is a complex system of separate processes. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 57-62).

Detterman, D. K. (1986). Qualitative integration: The last word? En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 163-166).

Díaz, C. (1997). *Manual de Historia de las religiones*. Bilbao: Desclée De Brouwer; 2004.

Dick, Ph. K. (2007). *Cuentos completos* (Vol. III). Barcelona: Minotauro.

Dreyfus, H. L. (1979). From micro-worlds to knowledge representation: AI at an impasse. En J. Haugeland (Ed.), *Mind design* (pp. 161-204).

Dreyfus, H. L. (1992). *What computers still can't do*. Cambridge: The MIT Press; 1994.

Einstein, A. (1955). *Mi visión del mundo*. Barcelona: Tusquets; 2002.

Estes, W. K. (1986). Where is intelligence? En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 63-68).

Eysenck, H. J. (1986). Is intelligence? R. J. Sternberg & D. K. Douglas (Eds.), *What is intelligence?* (pp. 69-72).

Eysenck, H. J. (1998). *Intelligence: A new look*. New Jersey: Transaction Publishers; 2000.

Falguera, J. L., & Martínez, C. (1999). *Lógica clásica de primer orden*. Madrid: Trotta.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109* (42), 17028-17033.

Ferrater, J. (1965a). Sistema. En J. Ferrater, *Diccionario de filosofía* (Vol. II, pp. 687-690). Buenos Aires: Editorial Sudamericana; 2000.

Ferrater, J. (1965b). Técnica. En J. Ferrater, *Diccionario de filosofía* (Vol. II, pp. 763-764). Buenos Aires: Editorial Sudamericana; 2000.

Ferrater, J. (1965c). Modelo. En J. Ferrater, *Diccionario de filosofía* (Vol. II, p. 216). Buenos Aires: Editorial Sudamericana; 2000.

Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn effect*. Cambridge: Cambridge University Press; 2009.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. En J. Haugeland (Ed.), *Mind design* (pp. 307-338).

Fodor, J. A., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. En J. Haugeland (Ed.), *Mind design II* (pp. 309-350).

Fraile, G. (2000). Hobbes. En G. Fraile, *Historia de la filosofía* (Vol. III, p. 724). Madrid: Biblioteca de Autores Cristianos.

Franklin, S. (1995). *Artificial minds*. Cambridge: The MIT Press.

Fuster, J. (1997). Redes de memoria. *Investigación y ciencia*, *250*, 74-83.

Gadamer, H. G. (1960). *Verdad y método* (Vol. I). Salamanca: Sígueme; 1977.

García, E. (1996). Inteligencia y metaconducta. *Revista de psicología general y aplicada*, *50* (3), 297-312.

García, E. (2001a). *Mente y cerebro*. Madrid: Síntesis.

García, E. (2001b). Creatividad, mente y cultura. *Educación, desarrollo y diversidad*, *3*, 11-30.

García, E. (2009). Aprendizaje y construcción del conocimiento. En C. López Alonso & M. Matesanz del Barrio (Eds.), *Las plataformas del aprendizaje. Del mito a la realidad* (pp. 21-44). Madrid: Biblioteca Nueva.

García, E. (2010). Desarrollo de la mente: Filogénesis, sociogénesis y ontogénesis. En M. Maceiras & L. Méndez (Eds.), *Ciencia e investigación en la sociedad actual*. Salamanca: Editorial San Esteban.

Gardner, H. (1985). *La nueva ciencia de la mente*. Barcelona: Paidós; 1988.

Gardner, H. (1986). The waning of intelligence tests. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 73-76).

Gardner, H. (1993). *Inteligencias múltiples: La teoría en la práctica*. Barcelona: Paidós; 2010.

Gardner, H. (1999). Inteligencias múltiples. *Inteligencia viva. Investigación y ciencia*, Temas 17. 3er trimestre, 14-19.

Gershenfeld, N. (1999). *Cuando las cosas empiecen a pensar*. Barcelona: Granica; 2000.

Gibson, W. F. (1984). *Neuromante*. Barcelona: Planeta; 2006.

Glaser, R. (1986). Intelligence as acquired proficiency. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 77-84).

Goleman, D. J. (1995). *Inteligencia emocional*. Barcelona: Kairós; 2004.

Goodnow, J. J. (1986). A social view of intelligence. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 85-90).

Guijarro, V., & González, L. (2010). *La quimera del autómata matemático*. Madrid: Cátedra.

Habermas, J. (1968a). *Conocimiento e interés*. Madrid: Taurus; 1982.

Habermas, J. (1968b). *Ciencia y técnica como ideología*. Madrid: Tecnos; 1986.

Hare, R. D. (1993). *Sin conciencia*. Barcelona: Paidós; 2012.

Hare, R. D., & Babiak, P. (2006). *Snakes in suits: When psychopaths go to work*. New York: HarperCollins; 2007.

Haugeland, J. (1978). The nature and plausibility of cognitivism. En J. Haugeland (Ed.), *Mind design* (pp. 243-281).

Haugeland, J. (1981a). Semantic engines. En J. Haugeland (Ed.), *Mind design* (pp. 1-34).

Haugeland, J. (Ed.) (1981b). *Mind design*. Cambridge: The MIT Press; 1985.

Haugeland, J. (1996). What is mind design? En J. Haugeland (Ed.), *Mind design II* (pp. 1-28).

Haugeland, J. (Ed.) (1997). *Mind design II*. Cambridge: The MIT Press.

Hawkins, J., & Blaskeslee, S. (2004). *Sobre la inteligencia*. Madrid: Espasa.

Heidegger, M. (1927). *El ser y el tiempo*. México: Fondo de Cultura Económica; 2008.

Heidegger, M. (1952). *Arte y poesía*. Madrid: Fondo de Cultura Económica; 1999.

Hobbes, Th. *Leviatán*. Madrid: Alianza; 2001.

Horkheimer, M. (1941). *Teoría crítica*. Madrid: Amorrortu; 2003.

Horkheimer, M. (1947). *Crítica de la razón instrumental*. Madrid: Trotta; 2002.

Horkheimer, M., & Adorno, Th. W. (1944). *Dialéctica de la Ilustración*. Madrid: Trotta; 2009.

Horn, J. L. (1986). Some thoughts about intelligence. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 91-96).

Humphreys, Ll. G. (1986). Describing the elephant. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 97-100).

Hunt, E. (1986). The heffalump of intelligence. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 101-108).

Isaacson, W. (2011). *Steve Jobs: La biografía*. Madrid: Debate.

Jackson, P. (1986). *Introduction to expert systems*. New York: Addison-Wesley; 1988.

Jensen, A. R. (1986). Intelligence: Definition, measurement and future research. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 109-112).

Kaku, M. (2011). *La física del futuro*. Barcelona: Debate.

Kandel, E. R., Schwartz, J. H., & Jessell, Th. M. (Eds.) (1995). *Essentials of neural science and behavior*. Stamford: Prentice Hall.

Kaufman, J. C., Kaufman, S. B., & Plucker, J. A. (2011). Contemporary theories of intelligence. En D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology*.

Kuhn, Th. S. (1970). *The structure of scientific revolutions*. Chicago: The University of Chicago Press; 1996.

Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence*. New York: Penguin Books; 2000.

Lakatos, I., & Musgrave, A. (Eds.) (1965). *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press; 1970.

Marcuse, H. (1941). *Razón y revolución*. Barcelona: Altaya; 1994.

Marcuse, H. (1955). *Eros y civilización*. Barcelona: Planeta; 1985.

Marcuse, H. (1958). *El marxismo soviético*. Madrid: Alianza; 1969.

Marcuse, H. (1964). *El hombre unidimensional*. Barcelona: Orbis; 1984.

Marcuse, H. (1967). *El final de la utopía*. Barcelona: Planeta; 1986.

Marr, D. C. (1977). Artificial intelligence: A personal view. En J. Haugeland (Ed.), *Mind design* (pp. 129-142).

Marx, K., & Engels, F. *Manifiesto del partido comunista*. Madrid: Biblioteca Nueva; 2007.

McCarthy, J., & Hayes, P. J. Some philosophical problems from the standpoint of artificial intelligence. Descargado de www-formal.stanford.edu/jmc.

McCorduck, P. (1979). *Machines who think*. San Francisco: W.H. Freeman & Company; 1981.

McDermott, D. V. (1976). Artificial intelligence meets natural stupidity. En J. Haugeland (Ed.), *Mind design* (pp. 143-160).

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: The University of Chicago Press; 1974.

Michie, D. (1974). *On machine intelligence*. Edinburgh: Edinburgh University Press.

Minsky, M. L. (1975). A framework for representing knowledge. En J. Haugeland (Ed.), *Mind design* (pp. 95-128).

Minsky, M. L. (1985). *The society of mind*. New York: Touchstone; 1988.

Newell, A., & Simon, H. A. (1975). Computer science as empirical enquiry: Symbols and search. En J. Haugeland (Ed.), *Mind design* (pp. 35-66).

Nietzsche, F. *La gaya ciencia*. Madrid: Akal; 2001.

Nietzsche, F. *Sobre verdad y mentira en sentido extramoral*. Valencia: Tilde; 2000.

Nietzsche, F. *Más allá del bien y del mal & Ecce homo*. Madrid: Libsa; 2000.

Ortega y Gasset, J. (1914). *Meditaciones del Quijote*. Madrid: Calpe; 1921.

Ortega y Gasset, J. (1930). *La rebelión de las masas*. Barcelona: Planeta; 1984.

Orwell, G. (1949). *1984*. Barcelona: Destino; 2005.

Packard, V. O. (1960). *The waste makers*. Harmondsworth: Penguin Books; 1966.

Pellegrino, J. (1986). Intelligence: The interaction of culture and cognitive processes. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 113-116).

Platón. *Fedro*. Madrid: Alianza; 2000.

Platón. *Obras completas*. Madrid: Aguilar; 1981.

Pohl, F. G. (1977). *Pórtico*. Barcelona: Planeta; 2005.

Popper, K. R. (1934). *The logic of scientific discovery*. London: Hutchinson; 1986.

Putnam, H. W. (1973a). Reductionism and the nature of psychology. En J. Haugeland (Ed.), *Mind design* (pp. 205-219).

Putnam, H. W. (1973b). Meaning and reference. *The journal of philosophy*, *70* (19), 699-711.

Pylyshyn, Z. (1974). Complexity and the study of artificial and Human Intelligence. En J. Haugeland (Ed.), *Mind design* (pp. 67-94).

Ramsey, W., Stich, S., & Garon, J. (1990). Connectionism, eliminativism and the future of folk psychology. En J. Haugeland (Ed.), *Mind design II* (pp. 351-376).

Reisberg, D. (Ed.) (2012). *The Oxford handbook of cognitive psychology*. New York: Oxford University Press.

Rich, E., & Knight, K. (1991). *Artificial intelligence*. Singapore: McGraw-Hill.

Rivadulla, A. (2003). Inconmensurabilidad y relatividad: Una revisión de la tesis de Thomas Kuhn. *Revista de Filosofía*, *28* (2), 237-259.

Rivadulla, A. (2009). El mito del método y las estrategias del descubrimiento científico: Inducción, abducción, preducción. En O. Pombo & A. Nepomuceno (Eds.), *Lógica e Filosofía da Ciência*. Centro de Filosofía da Ciências da Universidade de Lisboa, Colecçao Documenta 2, Lisboa, 231-246.

Rivadulla, A. (2010). Estrategias del descubrimiento científico: Abducción y preducción. En *Filosofía e História da Ciência no Cone Sul, 6º encontro*, 120-129.

Rivera, J. A. (2003). *Lo que Sócrates diría a Woody Allen*. Madrid: Espasa; 2005.

Robinet, A. (1973). *Mitología, filosofía y cibernética*. Madrid: Tecnos; 1982.

Rodríguez, R. (1987). *Heidegger y la crisis de la época moderna*. Madrid: Síntesis; 2006.

Rorty, R. M. (1979). *Philosophy and the mirror of Nature*. New Jersey: Princeton University Press; 1980.

Rosenberg, J. F. (1990). Connectionism and cognition. En J. Haugeland (Ed.), *Mind design II* (pp. 292-308).

Rumelhart, D. E., McClelland, J. L., & The PDP Research Group (1986). *Parallel distributed processing* (Vol. I). Cambridge: The MIT Press; 1989.

Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. En J. Haugeland (Ed.), *Mind design II* (pp. 205-232).

Russell, B. (1918). *The philosophy of logical atomism*. London: Routledge; 2010.

Santos, J., & Duro, R. J. (2005). *Evolución artificial y robótica*. Madrid: Ra-Ma.

Salfellner, H. (Ed.) (2011). *El golem de Praga: Leyendas judías del gueto*. EU: Vitalis.

Scarr, S. W. (1986). Intelligence: Revisited. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 117-120).

Schank, R. C. (1986). Explaining intelligence. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 121-132).

Schank, R. C. (1999). *Dynamic memory revisited*. Cambridge: Cambridge University Press.

Searle, J. R. (1980). Minds, brains and programs. En J. Haugeland (Ed.), *Mind design* (pp. 282-306).

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.

Searle, J. R. (1984). *Minds, brains and science*. Cambridge: Harvard University Press.

Searle, J. R. (1990). ¿Es la mente un programa informático? *Investigación y ciencia*, *162*, 10-16.

Searle, J. R. (1996). Dos biólogos y un físico en busca del alma. *Mundo Científico*, *170*, 654-669.

Shepherd, G. M. (1988). *Neurobiology*. New York: Oxford University Press.

Simon, H. A. (1981). *The sciences of the artificial*. Cambridge: The MIT Press; 1996.

Smolensky, P. (1989). Connectionist modeling: Neural computation, mental connections. En J. haugeland (Ed.), *Mind design II* (pp. 233-250).

Snow, R. E. (1986). On intelligence. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 133-140).

Spengler, O. (1931). *El hombre y la técnica*. Buenos Aires: Editorial Ver; 1963.

Spinoza, B. *Ética demostrada según el órden geométrico*. México: Fondo de Cultura Económica; 1996.

Sternberg, R. J. (1986). Intelligence is mental self-government. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 141-148).

Sternberg, R. J., & Berg, C. A. (1986). Quantitative integration: Definitions of intelligence: A comparison of the 1921 and 1986 symposia. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 155-162).

Sternberg, R. J., & Detterman, D. K. (Eds.) (1986). *What is intelligence?* New Jersey: Ablex Publishing.

Sternberg, R. J. (1999). Intelligence. En R. Wilson & F. Keil (Eds.) (1999), *The MIT encyclopedia of the cognitive sciences*. Cambridge: The MIT Press; 1999.

Sternberg, R. J. (Ed.) (2011). *The Cambridge handbook of intelligence*. Cambridge: Cambridge University Press.

Tammet, D. P. (2006). *Born on a blue day*. London: Hodder & Stoughton.

Taylor, F. W. (1911). *The principles of scientific management*. New York: Dover Publications; 1998.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. En J. copeland (Ed.), *The essential Turing* (pp. 58-90).

Turing, A. M. (1948). Intelligent machinery. En J. Copeland (Ed.), *The essential Turing* (pp. 410-432).

Turing, A. M. (1950). *¿Puede pensar una máquina?* Buenos Aires: Almagesto; 1990.

Turing, A. M. (1951). Intelligent machinery, a heretical theory. En J. Copeland (Ed.), *The essential Turing* (pp. 472-475).

Turing, A. M., Braithwaite, R., Jefferson, G., & Newman, M. (1952). Can automatic calculating machines be said to think? En J. Copeland (Ed.), *The essential Turing* (pp. 494-506).

van Gelder, T. (1996). Dynamics and cognition. En J. Haugeland (Ed.), *Mind design II* (pp. 421-450).

von Neumann, J. (1956). *The computer and the brain*. London: Yale University Press; 2000.

Weizenbaum, J. (1976). *La frontera entre el ordenador y la mente*. Madrid: Pirámide; 1978.

Winner, L. (1986). *La ballena y el reactor*. Barcelona: Gedisa; 2008.

Winston, P. H. (1981). *Artificial intelligence*. New York: Addison-Wesley; 1984.

Wittgenstein, L. (1921). *Tractatus logico-philosophicus*. Madrid: Alianza; 2000.

Wittgenstein, L. (1935). *The blue and brown books*. New York: Harper & Row; 1965.

Wittgenstein, L. (1951). *Sobre la certeza*. Barcelona: Gedisa; 2003.

Wittgenstein, L. (1953). *Investigaciones filosófi*cas. Barcelona: Crítica; 2012.

Woolgar, S. (1988). *Ciencia: Abriendo la caja negra*. Barcelona: Anthropos; 1991.

Worchel, S., Cooper, J., Goethals, G. R., & Olson, J. M. (2000). *Psicología social*. Madrid: Thomson; 2003.

Zigler, E. F. (1986). A developmental approach. En R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 149-152).