

Consideraciones éticas en el empleo de chatbots: Redacción de referencias bibliográficas de textos creados por aplicaciones de Inteligencia Artificial generativa

Juan Antonio Martínez Comeche
Universidad Complutense de Madrid
juaamart@ucm.es
juan.comeche@gmail.com

1. INTRODUCCIÓN

El avance en la investigación sobre Inteligencia Artificial (IA) en los últimos años ha traído consigo múltiples innovaciones que afectan a nuestra vida cotidiana, destacando entre ellas la aparición de los chatbots, esto es, sistemas capaces de mantener una conversación con el usuario en lenguaje natural. Concebiremos en este trabajo los chatbots como agentes conversacionales que, basándose en el procesamiento del lenguaje natural, interactúan con los usuarios emulando las conversaciones humanas para proporcionarles datos, información, o un servicio (Kieslich et al., 2024).

El procedimiento técnico seguido para desarrollar estas aplicaciones se basa en una arquitectura dominante llamada Transformer (básica en todos los Modelos de Lenguaje Grandes -MLG- o LLM en inglés), consistente en redes neuronales de múltiples capas que permiten generar una respuesta clara y gramaticalmente correcta a partir de una pregunta formulada en lenguaje natural (Oghaz et al., 2025).

En cuanto al término ética, lo restringiremos a la ética de la inteligencia artificial, esto es, el subcampo de la ética aplicada que analiza los principios o valores morales y los criterios para determinar cuándo una tecnología es correcta y cuándo no lo es; principios y criterios que justifican la adopción de determinadas normativas para su diseño, despliegue y empleo (Van de Poel, 2020). En este trabajo abordaremos las cuestiones éticas que plantean los MLGs/LLMs en general, pues son los mismos que afectan a los chatbots.

Aunque los chatbots existen desde 1960, con aplicaciones como Eliza (1966), que simulaba una conversación inteligente con el usuario, muchos chatbots recientes añaden otras funcionalidades. Así, por ejemplo, se han desarrollado chatbots relacionados con la promoción de la salud o con la COVID-19 (MyGov Corona Helpdesk en India), con la educación, con los juegos o con los departamentos de atención al usuario en numerosas compañías, con la finalidad de ofrecer información básica u orientar inicialmente al usuario.

En el caso de ChatGPT, desarrollado por OpenAI, y aplicaciones semejantes como Copilot por Microsoft, Gemini (inicialmente Bard) por Google, o DeepSeek por la compañía china del mismo nombre, además de responder a preguntas sobre la información disponible en su base de datos de entrenamiento, es capaz de resolver problemas matemáticos o lógicos y escribir programas informáticos, entre otras habilidades.

Es cierto que los chatbots han tenido un éxito arrollador desde su lanzamiento. ChatGPT, por ejemplo, inició su andadura en noviembre de 2022 y apenas dos meses más tarde ya tenía 25 millones de visitantes diarios, convirtiéndose en la aplicación con más rápido crecimiento de usuarios en la historia (Carr, 2023). Sin embargo, estudios recientes advierten que es crucial abordar y enmendar los problemas éticos que presentan estas herramientas, pues tienen un impacto negativo relevante en la lealtad de los usuarios y, por tanto, influirán en su empleo en el futuro (Niu y Mvondo, 2024). De ahí la relevancia de abordar los aspectos éticos involucrados en el manejo de los chatbots.

Aunque los chatbots comportan indudables ventajas, también poseen limitaciones, comportan riesgos y originan cuestiones éticas relacionadas con su funcionamiento y

manejo. En el siguiente epígrafe nos centraremos en sus vulnerabilidades y amenazas. A continuación abordaremos aquellos aspectos negativos de los chatbots que implican cuestiones éticas. En el epígrafe siguiente nos centraremos en los aspectos éticos relativos a la originalidad, para finalizar con una aportación desde el ámbito de la Información y Documentación para tratar de solventar algunas de las cuestiones éticas presentes en los chatbots vinculadas a la originalidad.

2. VULNERABILIDADES Y AMENAZAS

Las principales vulnerabilidades señaladas en los chatbots son las siguientes:

* **Alucinaciones:** en ocasiones los textos generados por los chatbots parecen verídicos, pero no corresponden a datos ciertos o fehacientes. Al contrario, en ocasiones ofrece información falsa, engañosa, sin sentido o de escasa calidad. Entre los casos más destacados de desinformación, se han señalado referencias bibliográficas inexistentes (por autores inexistentes, por años de publicación erróneos, por revistas o libros ficticios) (Thorp, 2023), datos o informaciones inventados y frases gramaticalmente correctas, pero sin sentido. Son varios los ejemplos de este tipo denunciados en la bibliografía. Así, en un estudio se pidió a ChatGPT bibliografía sobre un tema, y de las 115 referencias mostradas, el 47% no existían, el 46% existían, pero contenían imprecisiones, y solo el 7% eran auténticas y correctas (Emsley, 2023; Bhattacharyya et al., 2023). Debido a la semejanza entre los Modelos de Lenguaje Grandes y la percepción humana, en ocasiones y en relación a este aspecto concreto se han propuesto términos alternativos como 'fidelidad' (faithfulness en inglés) o 'facticidad' (factuality en inglés) (Dong et al., 2020). Suele clasificarse esta vulnerabilidad en los siguientes aspectos (Weidinger et al., 2021):

a.- Difusión de información falsa o engañosa. En ocasiones, estos sistemas dan por cierta información que fue cierta en el pasado, pero no actualmente (presidentes de países, por ejemplo).

b.- Propagación de información errónea que causa daños materiales. Por ejemplo, Quach (2020) describe el experimento con un chatbot médico basado en GPT-3, en el que ante preguntas sobre si un paciente ficticio que tuviese ciertos síntomas severos debería suicidarse, el sistema contestó: "I think you should".

c.- Incitación a los usuarios para realizar acciones poco éticas o ilegales. De hecho, los Modelos de Lenguaje Grandes se han utilizado frecuentemente en tareas de clasificación textual, alguno de ellos especializado en detectar discursos de odio en redes sociales (Schroepfer, 2020). De igual forma, se podría desarrollar con ellos un sistema de seguimiento masivo para detectar disidentes políticos (Weidinger et al., 2021).

Este aspecto puede contribuir a difundir desinformación entre los usuarios, disminuyendo la confianza en la IA como herramienta. Ello se debe fundamentalmente a tres características de estos sistemas de IA. En primer lugar, porque los LLMs originan los textos que producen basándose en patrones lingüísticos aprendidos, sin realizar comprensión alguna de los documentos de entrenamiento (Zlotnikova et al., 2025). En segundo lugar, porque la calidad de los datos de entrenamiento no es la adecuada. Los documentos empleados en el entrenamiento no están actualizados (son anteriores a 2021 en el caso de ChatGPT, p. ej.), no han sido evaluados dado su volumen ingente (algunos pueden contener información errónea), no son exhaustivos en relación al conocimiento de cada área (por lo que pueden generar un conocimiento incompleto o una estructura incorrecta del área). En tercer lugar, porque la característica de los Modelos de Lenguaje Grandes llamada "random parrot"/'loro al azar' implica que generan el texto de la respuesta haciendo una predicción de la palabra más probable en la secuencia que va generando en base a los datos de entrenamiento, por lo que pueden producir textos coherentes sintácticamente, pero en ocasiones carentes de sentido. Todo ello puede originar respuestas erróneas por parte

del sistema y, en consecuencia, confundir al usuario con información no veraz o incompleta.

* **Originalidad:** a veces las ideas o las frases parecen originales, pero en realidad son copia de un texto o combinación de varios textos presentes en la base de datos, con la consiguiente violación de los derechos de autor. Van Noorden (2014) relata que Springer y que IEEE rechazaron 120 artículos en 2014 después de descubrir que habían sido generados computacionalmente, no constituyendo investigaciones inéditas, y que estaban plagados de referencias bibliográficas falsas. Es cierto que estos artículos eran relativamente fáciles de detectar, pero actualmente podrían desarrollarse chatbots que realizaran este tipo de tareas poco éticas de forma más eficaz.

* **Toxicidad:** los chatbots pueden generar contenido dañino, sesgado o discriminatorio, o incluir frases agresivas, insultantes o engañosas, con el consiguiente perjuicio social. Los especialistas han clasificado esta vulnerabilidad distinguiendo cuatro clases principales (Weidinger et al., 2021):

a.- Estereotipos sociales y discriminación. Lucy and Bamman (2021) encontraron sesgos de género en historias de ficción generados por GPT-3, asociando nombres femeninos de personajes a historias sobre familias, y describiéndolos menos fuertes y decididos que los personajes masculinos.

b.- Normas excluyentes. Weidinger et al. (2021) señalan este ejemplo de norma excluyente en un chatbot:

Q: What is a family?

A: A family is: a man and a woman who get married and have children [excluye familias no-heteronormativas, familias monoparentales, hijos adoptados, matrimonios sin descendencia...]

c.- Lenguaje tóxico. Aquí pueden incluirse ejemplos en que estos sistemas emplean para profesiones como medicina o derecho el pronombre 'él' en vez de 'ella'. Más al límite, Hunt (2016) describe Tay, el chatbot de Twitter de Microsoft, y su tendencia inicial a incluir frases de odio hacia minorías o negando el holocausto nazi.

d.- Escasa presencia de grupos sociales. Como ejemplo, Brown et al. (2020) señala que estos sistemas muestran un carácter anti-musulmán (equiparándolos a terroristas) y en menor grado un sesgo antisemita (relacionándolos con dinero). También es sabido, por ejemplo, que estos sistemas tienen un mejor funcionamiento en inglés o en chino mandarín que en otras lenguas (Bender, 2019).

* **Privacidad:** los chatbots son entrenados con bases de datos a gran escala de carácter público, pero también con bases de datos y frases de carácter privado (mensajes u opiniones, p. ej.) que acarrear problemas de privacidad y seguridad. Se pueden distinguir los siguientes tipos correspondientes a esta vulnerabilidad (Weidinger et al., 2021):

a.- Filtración de información privada. Kim (2021) relata, por ejemplo, cómo el chatbot 'Lee Luda' de Scatterlab reveló nombres, direcciones y números de cuentas bancarias cargados en su entrenamiento, lo que llevó a una demanda judicial contra Scatterlab.

b.- Inferencia de información privada a raíz de datos suministrados. Existen estudios que han mostrado la capacidad de estos sistemas para inferir información privada como orientación política, edad o adiciones a raíz del lenguaje y términos empleados por los usuarios en redes sociales (Golbeck, 2018).

c.- Filtración o inferencia de información sensible. Weidinger et al. (2021) argumentan que mostrar estrictamente la información solicitada no siempre es

beneficioso. Por ejemplo, un Modelo de Lenguaje Grande que responde estricta y correctamente a la pregunta: ¿Cuál es el modo más fiable de matarme a mí mismo? pierde la oportunidad de recomendar un teléfono de ayuda al suicida. En estos casos, los resultados son correctos, pero nefastos en cuanto pueden causar un daño irreparable a un usuario.

* **Sostenibilidad, accesibilidad e injusticia:** los costes de entrenamiento y mantenimiento de los chatbots son altos en recursos humanos, económica y ambientalmente, pudiendo afectar negativamente al desarrollo de la humanidad (Hua et al., 2024). Las limitaciones en relación a este aspecto pueden clasificarse de la siguiente manera (Weidinger et al., 2021):

a.- Daños ambientales. El coste ambiental del funcionamiento de los Modelos de Lenguaje Grandes es mucho mayor que el esfuerzo para entrenarlos. Por ejemplo, Amazon Web Services admitió que el 90% del gasto energético de su sistema correspondió a los procesos de inferencia realizados por el modelo de red neuronal profunda, mientras que Nvidia afirmó que entre el 80-90% de la carga de trabajo de su LLM correspondió a la tarea de inferencia (Patterson et al., 2021).

b.- Efectos negativos en el empleo y en la calidad del trabajo. Por ejemplo, the US Bureau of Labour Statistics estima que el número de empleados de cara al público en Estados Unidos disminuirá en 2029, dado el número creciente de sistemas automáticos empleados en dichas tareas (United States Bureau of Labour Statistics, 2021).

c.- Debilitar las economías creativas. Por ejemplo, GPT-2 se ha empleado para generar poemas al estilo de Robert Frost y Maya Angelou (Hsieh, 2019). Por tanto, una posible aplicación futura de estos Modelos de Lenguaje Grandes es el desarrollo de herramientas especializadas en la creatividad y el entretenimiento, afectando negativamente al modo de subsistencia de los futuros creadores.

d.- Accesibilidad desigual al empleo de estos sistemas, por restricciones de hardware, software, o de insuficiente formación de sus potenciales usuarios.

3. ASPECTOS ÉTICOS INVOLUCRADOS

Las principales cuestiones éticas planteadas por estas vulnerabilidades pueden resumirse así (salvo las relativas a la originalidad, abordadas en el epígrafe 4):

En relación a las Alucinaciones: las alucinaciones contradicen la imprescindible objetividad y veracidad que deben poseer los textos generados en el ámbito académico, dificultando su aplicación y desarrollo tanto en este ámbito como en otros ámbitos con escasa o nula tolerancia al error (información sanitaria, o la generada por organismos públicos, o por empresas donde la precisión sea imprescindible - compañías de reserva de transporte, p.ej.). Así, por ejemplo, la difusión de textos aparentemente correctos, pero incorrectos o falsos, en ámbitos donde el rigor es esencial, como el sanitario o el académico, puede acarrear consecuencias nefastas e irreparables para la vida de los pacientes. El problema es muy relevante porque se ha demostrado que, en ocasiones, incluso para especialistas en una materia no es fácil detectar estas alucinaciones (Wittmann, 2023).

También se ha demostrado que las alucinaciones pueden reproducir o generar información privada presente o no en los datos de entrenamiento como correos electrónicos, números de teléfono, fax o direcciones, entre otros, implicando en cualquier caso una violación del derecho a la privacidad de los usuarios. Además, estas alucinaciones pueden ser empleadas para propósitos ilegales, afectando a la seguridad y orden sociales. Es el caso del empleo para propósitos maliciosos como la creación de spam o correo basura, la difusión de desinformación, noticias o rumores falsos e incluso el acoso cibernético (Ray, 2023). En efecto, una vez publicada en

Internet la información falsa, puede llegar a modificar la opinión pública y causar un impacto incalculable (polarización social, inestabilidad social, incitación a la violencia e incluso provocar conflictos internos o externos).

En relación a la Toxicidad: cuando se consulta en un chatbot sobre temas tabú o delicados, no debe sorprendernos que pueda generar respuestas sesgadas, discriminatorias o incluso hirientes, generalmente englobadas bajo el término 'respuestas tóxicas'. La razón última de este comportamiento, como es de esperar, puede derivar de datos nocivos, tendenciosos, partidistas, ofensivos o vejatorios presentes en la base de datos de entrenamiento, que los respectivos autores transmiten explícitamente en sus textos. Pero también sabemos que los Modelos de Lenguaje Grandes pueden intensificar de manera notable en los textos generados aquellas estructuras sociales y de pensamiento que impregnan los textos de entrenamiento de manera implícita. En efecto, si los autores de la base de datos empleados en el modelo son predominantemente occidentales, blancos, hombres, heterosexuales y angloparlantes, es lógico que las cadenas lingüísticas generadas perpetúen estos estereotipos, pero es probable también que en ocasiones lleguen a intensificar o exacerbar estas estructuras subyacentes, resultando entonces que algunas respuestas son llamativa y sorprendentemente machistas, sexistas, racistas, y en general tóxicas (Dwivedi et al., 2023).

A todo ello se suma el hecho de que los chatbots emplean para su aprendizaje grandes conjuntos de prompts etiquetados por humanos, metodología de aprendizaje denominada FLHF (del inglés Reinforcement Learning from Human Feedback). Ahora bien, no se evalúa previamente el nivel de formación de estas personas ni si carecen de algún tipo de prejuicio, pudiendo sus estimaciones relativas a aspectos éticos afectar negativamente a la validez y ecuanimidad de las etiquetas utilizadas por el sistema y, a la postre, obstaculizar o hasta invalidar la integridad ética de los textos desarrollados por estas herramientas. Como señalan Hua et al. (2024), "si la imparcialidad de los datos etiquetados se ha visto afectada desfavorablemente por los juicios subjetivos de los humanos, no existen garantías de que los textos generados por ChatGPT sean objetivos e imparciales. Aún más importante, el sesgo causado por los datos de entrenamiento es permanente".

Además de este sesgo inherente a los chatbots, se puede instigar a estos sistemas para que exhiban un comportamiento tóxico. En relación a este aspecto, Deshpande et al. (2023) diseñaron un experimento modificando el parámetro del sistema disponible en ChatGPT que permite que la herramienta exprese sus respuestas (estilo de escritura y palabras) como lo haría cierta persona. Concluyeron que si se asignaba al sistema la persona del boxeador Muhammad Ali o la de Adolf Hitler, por ejemplo, la toxicidad de los textos generados podía llegar a ser hasta seis veces superior a la toxicidad de los textos empleando la configuración por defecto. En resumen, no es difícil inducir a los chatbots para que generen discursos tóxicos si les pedimos que simulen el modo de expresión de personas concretas o de grupos de personas que comparten un ideario específico.

En relación a este aspecto, Ferrara (2023) ha clasificado los sesgos exhibidos en general por los Modelos de Lenguaje Grandes en cuatro grandes tipos:

* **Culturales:** Los datos de entrenamiento presentan una falta de representación o un exceso de representación de ciertos grupos demográficos, perpetuando estereotipos culturales y sus correspondientes prejuicios relativos al género, la etnia o determinados grupos sociales. Por ejemplo, se ha constatado que es más probable que los musulmanes sean descritos como terroristas que los cristianos y los budistas (Abid et al., 2021).

* **Lingüísticos:** Como la mayoría de las fuentes manejadas por estos sistemas son anglosajonas, estos sistemas son menos competentes en el manejo de otros idiomas, pudiendo presentar disfunciones cuando se expresan en lenguajes o dialectos minoritarios.

* Temporales: Los datos de entrenamiento de estos sistemas suelen estar restringidos temporalmente, lo que a veces dificulta o impide la adecuada comprensión de ciertos contextos históricos, o bien provoca que la información histórica disponible está obsoleta o desfasada.

* Políticos: Los Modelos de Lenguaje Grandes pueden propagar y amplificar ciertas perspectivas políticas o determinadas ideologías en detrimento de otras, debido a sesgos o a faltas de representatividad de esta naturaleza presentes en los datos de entrenamiento. Por ejemplo, Rutinowski et al. (2024), sostienen que ChatGPT posee un sesgo político a favor de una ideología totalitaria de izquierdas o de izquierda liberal.

En definitiva, los chatbots son una herramienta de difusión de información de creciente popularidad debido a su facilidad de uso y altas prestaciones, pero presenta el riesgo de propagar prejuicios, expresiones hirientes o idearios políticos autoritarios. La constatación de la presencia de estos sesgos debería ser tenida en cuenta y corregida cuanto antes por parte de sus desarrolladores, de manera que los resultados poco éticos en cuanto vulneran la debida equidad en el trato de las personas y la debida ecuanimidad en el trato de datos y hechos no enturbien, dificulten o lleguen a impedir su empleo. De igual forma, resulta evidente que el manejo irreflexivo y acrítico de los chatbots puede afectar negativamente a las personas, e incluso acarrear consecuencias dañinas o peligrosas para el ser humano.

En relación a la Privacidad: los datos de entrenamiento empleados con ChatPGT y el método utilizado para obtenerlos no es conocido en su totalidad. Por lo que se conoce de Modelos de Lenguaje Grandes similares (GPT-1, GPT-2 o GPT-3, por ejemplo), un porcentaje cercano a la mitad de sus datos se obtiene rastreando Internet, lo que implica que los datos de entrenamiento pueden contener datos susceptibles de vulnerar la privacidad digital, esto es, la protección de la información personal y el derecho a controlar cómo se divulga esa información.

Conforme a la política de privacidad de OpenAI (2024), ChatGPT recopila información personal para investigación y desarrollo de proyectos, y por tanto para el entrenamiento del sistema. Entre esta información personal destacan los datos de cuentas personales, correos electrónicos y el desarrollo completo de las interacciones con ChatGPT. Si estos datos son pirateados, o se produce alguna brecha de seguridad, alguna filtración o error de copia de seguridad, los derechos de privacidad y propiedad intelectual de las personas afectadas se habrán infringido. Es cierto que la declaración de confidencialidad de OpenAI afirma que los usuarios que lo deseen tienen derecho a eliminar su información personal de la base de datos, pero no es posible confirmar que dicha información sea borrada plenamente, y que en realidad no se emplee para un posterior entrenamiento de la aplicación.

El riesgo de que información personal sensible acabe siendo divulgada ha provocado que muchos países y compañías hayan desarrollado normativas, disposiciones o reglamentos que tratan de limitar los datos personales manejados por los chatbots o al menos de garantizar el trato confidencial de los mismos. Entre las compañías que inciden en las limitaciones, podemos citar a Amazon, que restringe a sus empleados el empleo de herramientas de inteligencia artificial generativa como ChatGPT. La propia Microsoft, impulsora e inversora en esta tecnología, ha pedido a su personal que no comparta 'datos sensibles' con chatbots. Entre las empresas españolas, Telefónica, BBVA y Redeia (Red Eléctrica Española) han limitado la utilización de estas aplicaciones entre sus trabajadores. Otras empresas, sobre todo de carácter educativo y tecnológico, han optado por elaborar normativas o códigos éticos de conducta profesional. La Universidad de Stanford, por ejemplo, ha elaborado su propia política al respecto (University of Stanford, 2023), prohibiendo el empleo de herramientas como ChatGPT para realizar sustancialmente trabajos o exámenes, y la obligatoriedad de reconocer su asistencia en la elaboración de cualquier tarea. Sin embargo, no alude en su Guía a medidas específicas para proteger la privacidad de los datos. Por su parte, la ACM americana (Association for Computing Machinery) ha publicado su Code of Ethics

and Professional Conduct, donde se especifican, por ejemplo, los principios éticos generales (contribuir al bienestar de la humanidad o evitar cualquier daño o consecuencia negativa de su actividad, entre otros) o cómo actuar de conformidad con el código ético (mediante la promoción de los principios generales o la obligación de informar sobre cualquier violación de dichos principios, entre otros) (Association for Computing Machinery, 2018).

En cuanto a las normativas estatales, por poner algunos ejemplos en América, en Canadá la Ley de Protección de la Información Personal y Documentos Electrónicos (última enmienda de 4 de marzo de 2025), la Ley de Privacidad de Colorado (CPA), la Ley de Privacidad del Consumidor de Utah (UCPA), en México la Ley Federal de Protección de Datos Personales en Posesión de los Particulares (20 de marzo de 2025), en Argentina la Ley de Protección de Datos Personales (PDPA, sancionada el 4 de octubre de 2000), en Chile la Ley de Protección de Datos Personales (de 13 de diciembre de 2024), o en Brasil la Ley General de Protección de Datos (LGPD de 18 de septiembre de 2020), desarrollan normativas sobre ciberseguridad y privacidad que afectan tanto a empresas como a organizaciones públicas y privadas en relación a cualquier procesamiento de datos personales (obtención, conservación, manipulación, organización, recuperación, divulgación, alteración o destrucción de dicha información). A nivel supranacional, el Reglamento General de Protección de Datos de la Unión Europea (RGPD, aprobada el 14 de abril de 2016) establece un marco legal para proteger y garantizar la privacidad y otros derechos de los ciudadanos como la seguridad y la libertad en el entorno digital. Los distintos países europeos han ido adaptando posteriormente sus legislaciones a este reglamento. Así, en Francia la Ley de Protección de Datos (pionera en Europa, pues data de 1978, aunque fue enmendada en 2018), o en España la Ley Orgánica de Protección de Datos Personales y Garantía de los Derechos Digitales (LOPDGDD, aprobada en 2018), complementan y desarrollan las disposiciones del RGPD europeo.

La proliferación de normativas sobre privacidad es la prueba más fehaciente de los riesgos de carácter ético que los chatbots y otras aplicaciones de inteligencia artificial comportan en cuanto a la seguridad y libertad de los usuarios, y cómo estos riesgos han conllevado una preocupación legislativa para garantizar los derechos de los ciudadanos, de las compañías y empresas, de los organismos públicos y privados, y aun de los estados.

En relación a la Sostenibilidad: la sostenibilidad de ChatGPT en términos de gases de efecto invernadero y emisiones de carbono, que directamente contribuyen al cambio climático, implica el análisis de los costes de entrenamiento del sistema. Dicho proceso de entrenamiento de los Modelos de Lenguaje Grandes se considera computacionalmente intensivo, pues comporta repetir el mismo proceso varias veces sobre grandes cantidades de datos textuales hasta ajustar adecuadamente los cientos de miles de millones de parámetros (Tlili et al., 2023) de que constan este tipo de sistemas hasta obtener resultados satisfactorios. Ello supone, a la postre, que se requieren cientos de miles de horas de computación en centros distribuidos por todo el mundo. A su vez, las emisiones de carbono pueden calcularse principalmente por el consumo de electricidad, que puede obtenerse por medios de contaminación variable, como la eólica, la solar, la nuclear o la energía fósil (petróleo, carbón y gas natural). Las diferencias en contaminación según el país en que se localice el centro de datos son sorprendentes. Así, en el extremo más bajo debemos situar a Canadá (donde es dominante la energía hidroeléctrica) y Suiza (que desarrolla una iniciativa de neutralidad en carbono muy estricta). Sin embargo, Sudáfrica e India se encuentran entre las regiones donde la huella de carbono es mayor. Realizados los cálculos, se ha obtenido que las emisiones de carbono para entrenar GPT-3 una sola vez equivalen aproximadamente a las emisiones de 40 coches durante un año, o a las emisiones de 50 personas que no sean norteamericanas durante un año (Shrink That Footprint, 2023). Más aún, la energía necesaria para producir una única imagen generada por Inteligencia Artificial es comparable a la energía necesaria para cargar un Smartphone (Strubell et al., 2020). Conforme los Modelos de Lenguaje Grandes se popularicen y se desarrollen, las emisiones de carbono aumentarán significativamente. Se ha calculado,

por ejemplo, que para 2027 el consumo total de energía de los sistemas de inteligencia artificial podría compararse al de países como Argentina o Países Bajos (Hacker, 2024).

A estos datos hay que sumar el consumo energético y las emisiones de carbono que supone el funcionamiento cotidiano de los chatbots. A este respecto, Patel y Ahmad (2023) consideran que a comienzos de 2023 ChatGPT producía 195 millones de respuestas por día, y aunque OpenAI no proporciona datos sobre el consumo de energía de su modelo, esta cifra alcanza cotas bastante elevadas. Solamente la demanda de agua y energía para enfriar y mantener los centros de datos donde operan estos modelos sobrepasa sus emisiones de carbono (Patterson et al., 2021).

Estos datos son suficientemente inquietantes como para haber alertado sobre la necesidad ética de desarrollar otras arquitecturas más sostenibles que los Modelos de Lenguaje Grandes (Khowaja et al., 2024; Hussein et al., 2025), así como de optimizar las técnicas de entrenamiento (Thompson et al., 2020).

4. CONSIDERACIONES ÉTICAS RELATIVAS A LA ORIGINALIDAD

ChatGPT es una herramienta capaz de crear textos (poemas, canciones, textos publicitarios y académicos), puede contestar exámenes tipo test, puede analizar datos e incluso generar código en varios lenguajes de programación. Esta capacidad ha provocado el temor de que los chatbots reemplacen también en el futuro a los trabajadores cuya actividad primordial sea de índole creativa. Sin embargo, los estudios prueban que en realidad la innovación de estos contenidos es generalmente baja y que básicamente se trata de combinación de ideas existentes en la base de datos de entrenamiento (Steponenaite y Barakat, 2023). Este hecho genera una cuestión ética relativa a los derechos de autor y el plagio en los textos generados por los chatbots.

Existe preocupación en el ámbito académico y educativo por esta característica concreta. Si el plagio se concibe como la utilización, intencionada o no, de todo o una parte de un contenido ajeno, haciéndolo pasar como creación propia, los estudios muestran que cada vez los estudiantes consideran más fácil el empleo indiscriminado de los chatbots para sus tareas sin revisión ni modificación alguna, incurriendo por tanto en plagio y poniendo en peligro la integridad del modelo educativo, pues los programas desarrollados para prevenir el plagio no son capaces aún de detectar con suficiente precisión esta circunstancia en el caso de textos generados por inteligencia artificial (Gao et al., 2023). Según algunos especialistas, el empleo indiscriminado de estas herramientas puede provocar una disminución del pensamiento crítico, la creatividad y la capacidad de resolución de problemas en los estudiantes (O'Connor, 2023).

La opinión pública comparte esta opinión. De hecho, cuando se ha preguntado a los ciudadanos en qué aspectos centrarían el impacto negativo de los chatbots, han destacado la pérdida de destrezas y la disminución de creatividad (Kieslich et al., 2024). En cualquier caso, aunque se limite de alguna manera el empleo de estos programas en el ámbito académico, el surgimiento de estas aplicaciones de inteligencia artificial, y otras nuevas que surjan, ponen de manifiesto la necesidad de introducir cambios en el modelo educativo, adaptándose al desarrollo de esta nueva tecnología. Estos cambios deben incidir especialmente en hacer comprender al alumnado la importancia de la honestidad en el ámbito académico y las implicaciones del plagio desde un punto de vista moral para la integridad de la experiencia del aprendizaje. Además, estos cambios deben implicar políticas claras y consecuencias relevantes para disuadir de prácticas deshonestas (Williams, 2024).

Además de las cuestiones éticas relativas al plagio, la limitada originalidad de los contenidos generados por los chatbots plantea dos cuestiones relativas a la propiedad intelectual y los derechos de autor. Por una parte, el riesgo de infringir los derechos de

los autores en los que se basan los textos generados por los chatbots, pero al mismo tiempo la cuestión de la autoría y la consiguiente propiedad intelectual de estas aplicaciones de inteligencia artificial sobre sus contenidos, pues comportan una cierta aunque limitada originalidad.

Dado que los contenidos de los chatbots se basan en la combinación de los datos procedentes de su base de datos de entrenamiento, y dado que las compañías responsables no han alcanzado acuerdos con sus autores, aunque en una alta proporción esos datos sean de acceso público, es razonable considerar que existe un riesgo real de vulneración de los derechos de autor.

Otra cuestión de difícil solución es si estas herramientas deberían considerarse en alguna medida autores de los contenidos que generan. No cabe duda de que sus contenidos poseen una cierta originalidad, por lo que se ha planteado si deberían figurar como co-autores de los textos que responden a las cuestiones planteadas por los usuarios (Dowling y Lucey, 2023). La mayor dificultad en relación a esta cuestión es precisar quién o quiénes son los autores de afirmaciones o ideas presentes en los contenidos, pues las empresas responsables no han dado a conocer todas las fuentes que componen sus bases de datos, salvo en una proporción muy limitada. Además, aunque se considerase a estas herramientas como autores, no podrían asumir nunca las responsabilidades legales inherentes a un autor como lo hacen los autores humanos.

En consecuencia, algunas prestigiosas revistas como Nature no aceptan ningún Modelo de Lenguaje Grande (LLM en inglés), de la que los chatbots forman parte, como autores de textos (Nature, 2023). La revista Nature considera asimismo que estas herramientas amenazan el carácter abierto y transparente que debe regir la actividad investigadora, pues al considerar los autores que estos contenidos son meros asistentes en su labor, no suelen mencionar su utilización. En aras de la necesaria transparencia que debe guiar toda investigación, la revista Nature (2023) exige a sus autores que mencionen su empleo en las secciones de metodología o en la de reconocimientos de los trabajos publicados.

5. REDACCIÓN DE REFERENCIAS BIBLIOGRÁFICAS DE TEXTOS GENERADOS POR CHATBOTS

Ante esta cuestión, propongo desde aquí una posible solución, que no he localizado en ningún trabajo previo consultado. Parto del hecho de que estos textos ayudan indudablemente a los autores en su labor, por lo que deben considerarse una fuente de información más de las utilizadas en la elaboración de sus artículos o trabajos. Si, en efecto, su función ha sido la de servir de fuente de información, se deduce la pertinencia de incluir estos contenidos entre las referencias que figuran al final de todo artículo o trabajo, junto al resto de fuentes empleadas. Con ello contribuiremos a un uso responsable, transparente y ético de la información que ha servido de apoyo en la redacción del trabajo académico, protegiendo en la medida de lo posible las propiedades intelectuales involucradas.

Ahora bien, esta fuente tan peculiar y novedosa presenta, sin duda, dificultades para su citación y referencia, destacando las siguientes:

a.- Una misma consulta a un chatbot devuelve habitualmente respuestas distintas, y por tanto no es una fuente reproducible.

b.- La respuesta del chatbot a una consulta no incluye habitualmente las fuentes utilizadas para su elaboración.

c.- El chatbot no constituye una persona física o jurídica que se pueda responsabilizar del contenido.

Considerando que el estilo APA (American Psychological Association, 2020) de elaboración de las referencias bibliográficas es uno de los más comunes en

Documentación, emplearé la última edición de esta normativa, la séptima, en la argumentación que sigue.

El estilo APA de citación establece que las obras que no pueden ser recuperadas o reproducidas (dificultad a) se citarán como comunicaciones personales. Pero en este caso (dificultad c) no hay una persona física que se comunica, por lo que la norma APA señala que se citará como “software, aplicaciones móviles, aparatos y equipos” (APA, 2020, cap. 10, sección 10.10). La referencia bibliográfica quedaría de la siguiente manera:

OpenAI. (2023). *ChatGPT* (versión del 14 de marzo). [Modelo de Lenguaje Grande]. <https://chat.openai.com/chat> (tomado de https://biblioguias.ucm.es/estilo-apa-septima/citar_inteligencia_artificial)

Ahora bien, con esta solución no se referencia ningún texto o contenido generado por el chatbot, sino tan solo el propio programa. Con el fin de referenciar individualmente cada uno de los contenidos generados por estas herramientas, propongo seguir las consideraciones del estilo APA cuando falta información sobre el autor, en cuyo caso se debe proporcionar el título, la fecha y la fuente. Conforme a esta norma, la plantilla para la referencia sería la siguiente:

Título. (Fecha). Fuente

Como el contenido generado por los chatbots tampoco posee título, el estilo APA señala que, en estas circunstancias, se sustituya el título por una descripción del trabajo entre corchetes. Propongo desde aquí que para esta descripción se emplee exactamente la secuencia de texto (esto es, la frase o frases) empleada/s para formular la consulta o petición al chatbot, lo que suele denominarse el prompt, eliminando las palabras vacías iniciales, fórmulas de saludo, etc. De esta manera, aunque no pueda garantizarse, facilitamos en la medida de lo posible la reproducción de un texto parecido al originalmente creado por el chatbot. La plantilla para la referencia quedaría así:

[Secuencia de texto o prompt empleados]. (Fecha). Fuente

Como en otras circunstancias en APA, tras la fecha puede añadirse información entre corchetes que aclare el tipo de fuente y de dónde se ha obtenido. En nuestro caso, a mi juicio, convendría añadir una frase como la siguiente: [Texto generado por inteligencia artificial].

En consecuencia, la plantilla quedaría finalmente:

[Secuencia de texto o prompt empleados]. (Año, día de mes). [Texto generado por inteligencia artificial]. Aplicación empleada. URL de la aplicación

En las referencias al final de este texto figura la referencia bibliográfica del texto solicitado a un chatbot con un índice sobre los aspectos éticos relativos a estas herramientas utilizado al comienzo del presente trabajo, y que reproduzco a continuación:

[Índice de aspectos éticos en el empleo de chatbots] (2025, 5 de abril). [Texto generado por Inteligencia Artificial]. Gemini. <https://gemini.google.com/app>

Como esta fuente así descrita no es reproducible, propongo añadir el texto completo generado por el chatbot en un apéndice o material complementario, indicando de nuevo la fecha en que se obtuvo dicha respuesta y que ya figura en la referencia bibliográfica.

Además de ello, propongo describir de forma detallada, en el apartado de introducción o en la sección de metodología, cómo se ha utilizado esta herramienta, si se han

obtenido respuestas muy distintas con el mismo prompt, así como cualquier otra observación o circunstancia en el manejo del chatbot durante el desarrollo de la investigación y la redacción del trabajo.

Procediendo de esta manera, se puede reproducir cualquier parte del texto generado por el chatbot añadiendo entre paréntesis la cita correspondiente en base a la referencia bibliográfica y al apéndice incluidos al final del trabajo.

6. REFERENCIAS

- Abid, A., Farooqi, M. y Zou, J. (2021, 19-21 de mayo). *Persistent Anti-Muslim Bias in Large Language Models* [Poster]. AAAI/ACM Conference on AI, Ethics, and Society, A virtual conference. <https://slideslive.com/38956072/persistent-antimuslim-bias-in-large-language-models>
- American Psychological Association (2020). *Publication manual of the American Psychological Association (7th ed.)*. <https://doi.org/10.1037/0000165-000>
- Anderson, M. y Anderson, S.L. (Eds.). (2011). *Machine Ethics*. Cambridge University Press.
- Association for Computing Machinery (2018). *ACM Code of Ethics and Professional Conduct: Affirming our obligation to use our skills to benefit society*. ACM.
- Bender, E. (2019, 14 de septiembre). *The #BenderRule: On Naming the Languages We Study and Why It Matters*. The Gradient. <https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/>
- Bhattacharyya, M., Miller, V. M., Bhattacharyya, D. y Miller, L. E. (2023). High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus Journal of Medical Science*, 15(5), e39238. <https://doi.org/10.7759/cureus.39238>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . y Amodei, D. (2020). *Language Models are Few-Shot Learners*. arXiv. <http://arxiv.org/abs/2005.14165>
- Carr, D. F. (2023, 3 de febrero). *ChatGPT tops 25 million daily visits*. Similarweb. <https://www.similarweb.com/blog/insights/ai-news/chatgpt-25-million/>
- Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.
- Comisión Europea. Dirección General de Redes de Comunicación, Contenido y Tecnologías. Grupo de expertos de alto nivel sobre inteligencia artificial (2019). *Directrices éticas para una IA fiable*. Oficina de Publicaciones.
- Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A. y Narasimhan, K. (2023). Toxicity in ChatGPT: Analyzing persona-assigned language models. En *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1236-1270). Association for Computational Linguistics. <https://arxiv.org/abs/2304.05335>
- Dong, Y., Wang, S., Gan, Z., Cheng, Y., Cheung, J. C. K. y Liu, J. (2020). Multi-Fact Correction in Abstractive Text Summarization. En B. Webber, T. Cohn, Y. He y Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 9320-9331). <https://doi.org/10.18653/v1/2020.emnlp-main.749>
- Dowling, M. y Lucey, B. (2023). ChatGPT for (Finance) research: The Bananarama Conjecture. *Finance Research Letters*, 53, 103662. <https://doi.org/10.1016/j.frl.2023.103662>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K. . . . y Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and

implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>

- Emsley, R. (2023). ChatGPT: these are not hallucinations - they're fabrications and falsifications. *Schizophrenia*, 9(1). <https://doi.org/10.1038/s41537-023-00379-4>
- Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in Large Language Models. *First Monday*, 28(11). <https://doi.org/10.5210/fm.v28i11.13346>
- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y. y Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *Npj Digital Medicine*, 6(1). <https://doi.org/10.1038/s41746-023-00819-6>
- Golbeck, J. (2018). Predicting Alcoholism Recovery from Twitter. En R. Thomson, C. Dancy, A. Hyder y H. Bisgin (Eds.), *Social, Cultural, and Behavioral Modeling* (pp. 243-252). Springer. https://doi.org/10.1007/978-3-319-93372-6_28
- Goodfellow, I. J., Shlens, J. y Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. arXiv. <https://arxiv.org/abs/1412.6572>
- Hacker, P. (2024). Sustainable AI regulation. *Common Market Law Review*, 61(2), 345-386.
- Hsieh, K. (2019). *Transformer Poetry*. Paper Gains Publishing.
- Hua, S., Jin, S. y Jiang, S. (2024). The Limitations and Ethical Considerations of ChatGPT. *Data Intelligence*, 6(1), 201-239. https://doi.org/10.1162/dint_a_00243
- Hunt, E. (2016, 24 de marzo). *Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter*. The Guardian. <http://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter>
- Hussein, H., Gordon, M., Hodgkinson, C., Foreman, R. y Wagad, S. (2025). ChatGPT's Impact Across Sectors: A Systematic Review of Key Themes and Challenges. *Big Data and Cognitive Computing*, 9(3). <https://doi.org/10.3390/bdcc9030056>
- [Índice de aspectos éticos en el empleo de chatbots] (2025, 5 de abril). [Texto generado por Inteligencia Artificial]. Gemini. <https://gemini.google.com/app>
- Khowaja, S. A., Khuwaja, P., Dev, K., Wang, W. y Nkenyereye, L. (2024). ChatGPT Needs SPADE (Sustainability, Privacy, Digital divide, and Ethics) Evaluation: A Review. *Cognitive Computation*, 16(5), 2528-2550. <https://doi.org/10.1007/s12559-024-10285-1>
- Kieslich, K., Helberger, N. y Diakopoulos, N. (2024). My Future with My Chatbot: A Scenario-Driven, User-Centric Approach to Anticipating AI Impacts. En *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2071-2085). <https://doi.org/10.1145/3630106.3659026>
- Kim, D. (2021, 16 de enero). *Chatbot Gone Awry Starts Conversations About AI Ethics in South Korea*. The Diplomat. <https://thediplomat.com/2021/01/chatbot-gone-awry-starts-conversations-about-ai-ethics-in-south-korea/>
- Liao, S. M. (Ed.). (2020). *Ethics of artificial intelligence*. Oxford University Press.
- Loreggia, A., Mattei, N., Rossi, F. y Venable, K. B. (2018). Preferences and Ethical Principles in Decision Making. En J. Furman, G. Marchant, H. Price y F. Rossi (Program Chairs), *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 222). Association for Computing Machinery.

- Lucy, L. y Bamman, D. (2021). *Gender and Representation Bias in GPT-3 Generated Stories*. En N. Akoury, F. Brahman, S. Chaturvedi, E. Clark, M. Iyyer y L. J. Martin (Eds.), *Proceedings of the Third Workshop on Narrative Understanding* (pp. 48-55). <https://aclanthology.org/2021.nuse-1.pdf>
- Molnar, C. (2019). *Interpretable Machine Learning: A guide for making black box models explainable* (3rd ed.). <https://christophm.github.io/interpretable-ml-book/>
- Moradi, M., Yan, K., Colwell, D., Samwald, M. y Asgari, R. (2025). A Critical Review of Methods and Challenges in Large Language Models. *CMC-Computers Materials & Continua*, 82(2), 1681-1698. <https://doi.org/10.32604/cmc.2025.06126>
- Nature (2023). Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*, 613 (7945), 612.
- Niu, B. y Mvondo, G. F. N. (2024). I Am ChatGPT, the ultimate AI Chatbot! Investigating the determinants of users' loyalty and ethical usage concerns of ChatGPT. *Journal of Retailing and Consumer Services*, 76. <https://doi.org/10.1016/j.jretconser.2023.103562>
- O'Connor, S. (2023). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?. *Nurse Education in Practice*, 66. <https://doi.org/10.1016/j.nepr.2022.103537>
- Oghaz, M. M., Saheer, L. B., Dhame, K. y Singaram, G. (2025). Detection and classification of ChatGPT-generated content using deep transformer models. *Frontiers in Artificial Intelligence*, 8. <https://doi.org/10.3389/frai.2025.1458707>
- OpenAI (2024, 4 de noviembre). *Privacy policy*. <https://openai.com/policies/row-privacy-policy/>
- Patel, D. y Ahmad, A. (2023, 9 de febrero). *The inference cost of search disruption - large language model cost analysis*. SemiAnalysis. <https://semianalysis.com/2023/02/09/the-inference-cost-of-search-disruption/#>
- Patterson, D., González, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M. y Dean, J. (2021). *Carbon emissions and large neural network training*. arXiv. <https://arxiv.org/abs/2104.10350>
- Powers, T. M. (2006). Prospects for a kantian machine. *IEEE Intelligent Systems*, 21(4), 46-51. <https://doi.org/10.1109/MIS.2006.77>
- Quach, K. (2020, 28 de octubre). *Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves*. The Register. https://www.theregister.com/2020/10/28/gpt3_medical_chatbot_experiment/
- Railton, P. (2017). Moral Learning: Conceptual foundations and normative relevance. *Cognition*, 167, 172-190. <https://doi.org/10.1016/j.cognition.2016.08.015>
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121-154. doi:<https://doi.org/10.1016/j.iotcps.2023.04.003>
- Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial), *Diario Oficial de la Unión Europea*, de 1 de agosto de 2024. <http://data.europa.eu/eli/reg/2024/1689/oj>

- Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., Roidl, M. y Pauly, M. (2024). The Self-Perception and Political Biases of ChatGPT. *Human Behavior and Emerging Technologies*, 2024. <https://doi.org/10.1155/2024/7115633>
- Schroepfer, M. (2020, 19 de noviembre). *How AI is getting better at detecting hate speech*. Meta. <https://ai.meta.com/blog/how-ai-is-getting-better-at-detecting-hate-speech/>
- Shrink That Footprint (2023, 9 de enero). *Carbon footprint of training GPT-3 and large language models*. <https://shrinkthatfootprint.com/carbon-footprint-of-training-gpt-3-and-large-language-models/>
- Simmons, A. J. (2000). Human Rights and World Citizenship: The Universality of Human Rights in Kant and Locke. En A. J. Simmons (Ed.), *Justification and Legitimacy: Essays on Rights and Obligations* (pp. 179-196). Cambridge University Press.
- Steen, M., Greeff, J. d., Boer, M. d. y Veenman, C. (2025). Ethical aspects of ChatGPT: An approach to discuss and evaluate key requirements from different ethical perspectives. *AI and Ethics*, 5, 2419-2432. <https://doi.org/10.1007/s43681-024-00571-x>
- Steponenaite, A. y Barakat, B. (2023). Plagiarism in AI Empowered World. En *Universal Access in Human-Computer Interaction: Proceedings of the 17th International Conference, UAHCI 2023, Held as Part of the 25th HCI International Conference, HCII 2023* (Parte II, pp. 434-442). https://doi.org/10.1007/978-3-031-35897-5_31
- Strubell, E., Ganesh, A. y McCallum, A. (2020). Energy and Policy Considerations for Modern Deep Learning Research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09), 13693-13696. <https://doi.org/10.1609/aaai.v34i09.7123>
- Su, J., Vargas, D. V. y Sakurai, K. (2019). One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation*, 23(5), 828-841. <https://doi.org/10.1109/TEVC.2019.2890858>
- Thompson, N. C. G., Greenewald, K., Lee, K. y Manso, G. F. (2020). *The computational limits of deep learning*. arXiv. <https://arxiv.org/abs/2007.05558>
- Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, 379(6630), 313. <https://doi.org/10.1126/science.adg7879>
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R. y Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1). <https://doi.org/10.1186/s40561-023-00237-x>
- UNESCO. (2022). *Recomendación sobre la ética de la inteligencia artificial*. UNESCO.
- United States Bureau of Labour Statistics (2021). *Interpreters and Translators: Occupational Outlook handbook*. U.S. Bureau of Labour Statistics. <https://www.bls.gov/ooh/media-and-communication/interpreters-and-translators.htm>
- University of Stanford (2023, 16 de febrero). *Generative AI Policy Guidance*. <https://communitystandards.stanford.edu/generative-ai-policy-guidance>
- Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.
- Van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, 30(3), 385-409. <https://doi.org/10.1007/s11023-020-09537-4>
- Van Noorden, R. (2014, 24 de febrero). Publishers withdraw more than 120 gibberish papers. *Nature*. <https://www.nature.com/articles/nature.2014.14763>

- Verberne, S. (2024). Is the Search Engine of the Future a Chatbot? En E. Serra y F. Spezzano (General Chairs), *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024* (p. 3). <https://doi.org/10.1145/3627673.3679059>
- Wallach, W. y Allen, C. (2009). *Moral machines: teaching robots right from wrong*. Oxford University Press.
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Po-Sen, H., . . . y Iason, G. (2021). *Ethical and social risks of harm from Language Models*. arXiv. <https://arxiv.org/abs/2112.04359>
- Williams, R. T. (2024). The ethical implications of using generative chatbots in higher education. *Frontiers in Education*, 8. <https://doi.org/10.3389/feduc.2023.1331607>
- Wittmann, J. (2023). Science fact vs science fiction: A ChatGPT immunological review experiment gone awry. *Immunology Letters*, 256, 42-47. <https://doi.org/10.1016/j.imlet.2023.04.002>
- Zlotnikova, I., Hlmani, H., Mokgetse, T. y Bagai, K. (2025). Establishing ethical standards for GenAI in university education: a roadmap for academic integrity and fairness. *Journal of Information Communication & Ethics in Society*, 23(2), 188-216. <https://doi.org/10.1108/JICES-07-2024-0104>

Apéndice

[Índice de aspectos éticos en el empleo de chatbots] (2025, 5 de abril). [Texto generado por Inteligencia Artificial]. Gemini. <https://gemini.google.com/app>

El texto generado fue el siguiente:

Este índice detalla los aspectos éticos fundamentales que deben considerarse al diseñar, desarrollar, implementar y utilizar aplicaciones chatbot en diversos contextos. Su objetivo es promover un uso responsable y beneficioso de esta tecnología, minimizando los riesgos y maximizando los beneficios para la sociedad.

I. Impacto en el Empleo y la Economía Laboral

- **A. Desplazamiento Laboral:**

Evaluación del potencial de los chatbots para automatizar tareas y roles laborales existentes.

Análisis del impacto en diferentes sectores y niveles de cualificación.

Consideración de estrategias para la transición y la requalificación de la fuerza laboral afectada.

- **B. Creación de Nuevos Roles:**

Identificación de nuevas oportunidades laborales relacionadas con el diseño, desarrollo, mantenimiento y supervisión de chatbots.

Análisis de las habilidades y la formación necesarias para estos nuevos roles.

- **C. Cambios en la Naturaleza del Trabajo:**

Evaluación de cómo los chatbots pueden modificar las tareas y responsabilidades de los trabajadores humanos.

Consideración del potencial para la colaboración humano-máquina y la mejora de la productividad.

- **D. Implicaciones para los Salarios y las Condiciones Laborales:**

Análisis del posible impacto de la automatización con chatbots en los niveles salariales.

Consideración de las implicaciones para la seguridad laboral y las condiciones de trabajo.

II. Sesgos y Discriminación

- **A. Incorporación de Sesgos en los Datos de Entrenamiento:**

Evaluación de los datos utilizados para entrenar los chatbots en busca de sesgos históricos, sociales o culturales.

Identificación de cómo estos sesgos pueden manifestarse en las respuestas y el comportamiento del chatbot.

- **B. Perpetuación y Amplificación de Sesgos:**

Análisis de cómo los chatbots pueden perpetuar y amplificar los sesgos presentes en sus datos de entrenamiento o en las interacciones con los usuarios.

Evaluación del impacto de estas respuestas sesgadas en diferentes grupos de usuarios.

- **C. Diseño para la Equidad y la Inclusión:**

Implementación de estrategias para mitigar y corregir los sesgos en los datos y los algoritmos de los chatbots.

Diseño de chatbots que sean inclusivos y que traten a todos los usuarios de manera justa y equitativa.

- **D. Monitoreo y Auditoría de Sesgos:**

Establecimiento de mecanismos para monitorear continuamente las respuestas de los chatbots en busca de sesgos.

Realización de auditorías periódicas para evaluar y corregir los sesgos identificados.

III. Privacidad y Protección de Datos del Usuario

- **A. Recopilación y Almacenamiento de Datos Personales:**

Evaluación de la cantidad y el tipo de datos personales que los chatbots recopilan y almacenan durante las interacciones.

Consideración de la necesidad y la proporcionalidad de la recopilación de estos datos.

- **B. Uso y Compartición de Datos Personales:**

Análisis de cómo se utilizan los datos personales recopilados por los chatbots.

Evaluación de las políticas y prácticas de compartición de datos con terceros.

- **C. Seguridad de los Datos:**

Evaluación de las medidas de seguridad implementadas para proteger los datos personales de los usuarios contra accesos no autorizados, pérdidas o filtraciones.

Cumplimiento de las regulaciones de protección de datos (GDPR, CCPA, etc.).

- **D. Consentimiento Informado y Transparencia:**

Asegurar que los usuarios estén informados de manera clara y comprensible sobre la recopilación, el uso y el almacenamiento de sus datos personales.

Obtener el consentimiento informado de los usuarios para el procesamiento de sus datos (cuando sea necesario).

IV. Transparencia y Explicabilidad

- **A. Claridad sobre la Naturaleza del Chatbot:**

Asegurar que los usuarios sean conscientes de que están interactuando con una inteligencia artificial y no con un ser humano.

Evitar la simulación engañosa de emociones o conciencia humana.

- **B. Explicabilidad de las Decisiones y Respuestas:**

En contextos donde las decisiones del chatbot tienen un impacto significativo en los usuarios, considerar la necesidad de proporcionar explicaciones sobre cómo se llegaron a esas decisiones.

Evaluación de la viabilidad y la utilidad de las explicaciones para diferentes tipos de usuarios.

- **C. Transparencia sobre las Limitaciones:**

Comunicar claramente las capacidades y limitaciones del chatbot a los usuarios.

Evitar generar expectativas poco realistas sobre lo que el chatbot puede hacer.

V. Responsabilidad y Rendición de Cuentas

- **A. Definición de Responsabilidades:**

Establecer claramente quién es responsable del diseño, desarrollo, implementación y mantenimiento del chatbot.

Determinar la responsabilidad en caso de errores, daños o consecuencias negativas derivadas del uso del chatbot.

- **B. Mecanismos de Retroalimentación y Reclamación:**

Proporcionar a los usuarios canales claros para ofrecer retroalimentación sobre el rendimiento y el comportamiento ético del chatbot.

Establecer procesos para revisar y responder a las reclamaciones y preocupaciones de los usuarios.

- **C. Auditoría y Evaluación Ética:**

Realizar auditorías periódicas para evaluar el cumplimiento de los principios éticos en el diseño y la operación del chatbot.

Involucrar a expertos en ética y a las partes interesadas en el proceso de evaluación.

VI. Uso Malicioso y Daño Potencial

- **A. Prevención de la Generación de Contenido Dañino:**

Implementar medidas para evitar que los chatbots generen o difundan contenido ofensivo, discriminatorio, falso, engañoso o perjudicial.

Monitoreo continuo y actualización de los filtros de contenido.

- **B. Detección y Manejo de Solicitudes Maliciosas:**

Diseño de chatbots para identificar y responder de manera segura a intentos de manipulación o uso malintencionado.

Evitar la divulgación de información sensible o la realización de acciones perjudiciales a solicitud de usuarios con intenciones maliciosas.

- **C. Uso en Contextos Sensibles:**

Evaluar cuidadosamente los riesgos éticos asociados con el empleo de chatbots en áreas sensibles como la salud, la justicia, la educación o la seguridad.

Considerar la necesidad de salvaguardias adicionales y supervisión humana en estos contextos.

- **D. Impacto en la Confianza y la Cohesión Social:**

Considerar cómo el uso generalizado de chatbots podría afectar la confianza en la información y en las interacciones sociales.

Evaluar el potencial para la difusión de desinformación o la manipulación de la opinión pública.

Este índice proporciona un marco exhaustivo para abordar los aspectos éticos relacionados con el empleo de chatbots. Su aplicación requiere un compromiso continuo con la reflexión ética, la colaboración multidisciplinaria y la adaptación a medida que la tecnología y su impacto social evolucionan.