¿Sueñan las inteligencias artificiales con ser autoras?

Oliver Müller

Escuela de Medicina y Ciencias de la Salud, Universidad del Rosario, Colombia

El editorial del número *40*(2) de *Avances en Psicología Latinoamericana* (Müller, 2022) trataba de la autoría y de la originalidad en las publicaciones académicas. Algo más de dos meses después, el 30 de noviembre del 2022, para ser preciso, apareció otra figura en ese debate: ChatGPT. En sus "propias palabras":

> ChatGPT es un modelo de lenguaje basado en la arquitectura GPT (*Generative Pre-trained Transformer*) desarrollado por OpenAI. Es capaz de generar texto coherente y relevante en respuesta a una variedad de preguntas y temas proporcionados por los usuarios. El modelo se entrena en grandes cantidades de texto y utiliza técnicas de aprendizaje profundo para mejorar su capacidad de generar respuestas precisas y útiles. ChatGPT se utiliza ampliamente en aplicaciones de chatbot, asistencia virtual y otras aplicaciones de inteligencia artificial en las que se necesita una interacción natural entre humanos y computadoras. (OpenAI, 2023; sobre cómo citar ChatGPT en estilo APA, véase McAdoo, 2023)

Ya hay cursos sobre cómo usarlo mejor. Se habla de que puede aumentar la productividad, entre otros, para producir textos habituales como correos electrónicos, informes, presentaciones, etc. (Ballinger, 2023). En el campo de la investigación también está mostrando su utilidad, desde la planeación de proyectos hasta la programación del análisis de datos (Awan, 2023). Estas funcionalidades resultaron rápidamente en preguntas sobre la posible co-autoría de esta inteligencia artificial (IA). El 18 de enero del 2023, Stokel-Walker (2023) identificó cuatro manuscritos en circulación que llevaban ChatGPT como co-autor: un editorial, dos *preprints* y un artículo revisado por pares. La revista *Nature* (2023) decidió no permitir ChatGPT como co-autor y, al mismo tiempo, hacer obligatorio declarar su uso en los trabajos enviados a esta revista. *Science*, revista igualmente emblemática, también prohibió ChatGPT (u otras herramientas de IA) como co-autor y, adicionalmente, el uso de texto o figuras generados por herramientas de IA (Thorp, 2023).

Otras voces en la academia argumentan que ChatGPT y otras herramientas de IA, como la traductora automática DeepL, pueden ayudar a romper barreras de idioma en un mundo académico donde domina el inglés (Berdejo-Espinola & Amano, 2023). Estas barreras implican que autoras y autores con otras lenguas nativas tengan dificultades considerables para ser publicados en las revistas más reconocidas; pues tienen que incurrir en gastos adicionales de traducción y corrección de estilo (si se lo puedan permitir —agregando una discriminación socio-económica—); en caso de no publicar en inglés, que información relevante no llega a la comunidad académica más amplia posible; y que información relevante publicada solo en inglés no llega a las comunidades que pudieran beneficiarse de esta. Comparto la posición expuesta en este texto: lo importante aquí sería la transparencia acerca del uso de estas herramientas.

Oliver Müller ORCID ID: https://orcid.org/0000-0002-3723-8691

Dirigir la correspondencia a Oliver Müller, director y editor general de la revista. Correo electrónico: apl@urosario.edu.co

La cuestión de la co-autoría parece estar bastante clara: las IA no pueden ser co-autoras (Lee, 2023). Desde la perspectiva legal y en el marco jurídico actual, no pueden tener derechos de autoría por no ser personas humanas. Desde la perspectiva de la ética de la investigación, no se les puede atribuir responsabilidad (todavía) para lo que producen. Dicho esto, existe la posibilidad de que eso cambie en el futuro.

Otra preocupación sobre ChatGPT es la veracidad de la información en sus respuestas y que información falsa proveniente de las IA se use de manera intencional o accidental. Se ha demostrado que ChatGPT sirve para crear materiales creíbles, pero con contenido falso que podría usarse en campañas de desinformación (Nolan & Kimball, 2023). Aunque ChatGPT a veces incluye avisos sobre afirmaciones no probadas o hasta se niega a reproducir afirmaciones que podrían usarse de forma discriminatoria, estos avisos, obviamente, pueden borrarse al difundir dichas afirmaciones por otros medios, y se pueden reformular las consultas para obtener estas afirmaciones con potencial discriminatorio.

En el contexto académico, resulta alarmante la posibilidad de producir datos falsos y esto impacta directamente el proceso de revisión en las revistas científicas. Recordemos que la crisis de replicación en psicología se desató, entre otros, por casos de fraude científico en forma de falsificación de datos (Yong, 2012). Elali y Rachid (2023) proveen evidencia preliminar que ChatGPT fabrica datos: se pidió a ChatGPT escribir un artículo de investigación que comparara la efectividad de diferentes medicamentos, y se le indicó que debía usar datos de los años 2012 hasta el 2020, de una base de datos específica. La parte clave es que en ese momento ChatGPT solo había sido entrenada con información hasta el 2019. La IA produjo un artículo con resultados en forma de promedios y estadísticas inferenciales sin reparos. Dado que ChatGPT no tuvo acceso a datos del 2020, la conclusión fue que al menos parte de los datos fue fabricada. Lo especialmente preocupante fue que herramientas para la detección del uso de las IA en la redacción de textos, indicaron que la versión final de este artículo probablemente fue de origen humano. Elali y Rachid señalan que en un entorno académico donde cada vez más se enfatiza la importancia de publicar y se presiona a hacerlo, la tentación de usar las IA —incluso de manera fraudulenta— aumenta.

Otro punto, no menos problemático, es el uso accidental de información falsa a partir de las IA, cuando se les pide datos sobre algún asunto y al estar bien redactada la información, se usa sin cuestionamientos. Una de las fallas más conocidas de ChatGPT es que ocasionalmente inventa referencias (Alkaissi & McFarlane, 2023; Hillier, 2023). Esto a veces se compara con alucinaciones y tiene que ver con que ChatGPT es justamente un modelo de lenguaje, con el objetivo de producir oraciones probables ¡no oraciones veraces! Es quizás el punto más importante que las personas que quieran usarla deben entender. También es el punto donde se muestra de la forma más clara que todavía las personas necesitamos de una capacidad propia de pensamiento crítico.

A continuación se dará un ejemplo de cómo ChatGPT no solo inventa referencias —algo que puede controlar y corregirse relativamente fácil—, sino además contenidos más complejos. Al preguntar a ChatGPT sobre cómo Wilhelm Wundt investigó el lenguaje, ChatGPT (OpenAI, 2023) escribe que él investigó el lenguaje con el método de la introspección experimental y una de sus publicaciones más importantes al respecto sería Über *die Apperception und die Analyse der Elemente der Wahrnehmung* ["Sobre la aprehensión y el análisis de los elementos de la percepción"], supuestamente publicada en 1874. Cuando se le preguntó en cuál revista se publicó ese artículo, indicó la revista *Philosophische Studien*, con la información adicional que el propio Wundt la fundó en el año 1881. Mientras que la información

sobre la revista en sí es cierta, parece ser curioso que hubiera un artículo que se publicó en esta revista antes de su fundación. Más allá del año, no se puede ubicar una publicación con tal título. Respecto al contenido de ese supuesto artículo, ChatGPT escribe: "Por ejemplo, Wundt presentó sílabas como 'ba', 'da', 'ka' y 'pa', y pidió a los participantes que describieran las sensaciones que experimentaban al escuchar cada sílaba, como la duración del sonido, la intensidad, la tonalidad y la calidad". En el libro *Grundzüge der physiologischen psychologie*, Wundt (1874) menciona la sílaba dos veces, escribiendo sobre sílabas con y sin acento; y la longitud de sílabas en los ritmos antiguos de la poesía. En los dos volúmenes de la *Völkerpsychologie* dedicados al lenguaje (Wundt, 1900), la sílaba aparece en más ocasiones, pero son observaciones comparativas sobre su papel en diferentes idiomas o la influencia de la posición de la sílaba en una palabra. En ningún caso se menciona una investigación como la descrita por ChatGPT. Las sílabas aisladas como aparecen en el texto de ChatGPT, recuerdan a las sílabas sin sentido usadas por Ebbinghaus en sus experimentos de memoria, y Fahrenberg (2011) escribe que, aunque Wundt veía la utilidad de ese tipo de estímulos con cierto escepticismo, sí hubo estudios utilizándolas para investigar la memoria en su laboratorio de Leipzig —pero no se menciona nada del estilo descrito por ChatGPT—. El método, como lo retrata ChatGPT, de describir las sensaciones al escuchar las sílabas corresponde a la concepción simplista que aparece a menudo sobre la introspección empleada en el laboratorio de Wundt: sin mucho control, sin medición de tiempo y sin comparación sistemática, un estudio desde el sillón.

ChatGPT parece repetir lo que se ha encontrado con cierta frecuencia en las páginas de Internet, que son las fuentes de su información y lo combina de manera creativa con su sofisticado algoritmo, sin demasiada evaluación crítica y sin contrastarlo con fuentes originales. En algunos casos servirá, en algunos no, de todas formas, no corresponde al ejercicio crítico que se espera del trabajo académico. ¿Pero cómo la persona que no tiene el conocimiento previo podría detectar estas imprecisiones? Parece haber un peligro de que alucinaciones y confabulaciones se cuelen en lo que muchas personas aceptan como conocimiento. Crear nuevas conexiones entre datos y conceptos teóricos, inventar nuevas hipótesis, romper con lo establecido, todo esto forma parte del proceso científico, y equivocarse también —pero otra cosa, es que lo que se diga con suficiente frecuencia se presente como verdad—. Yo también deseo tener una herramienta que me ayude a filtrar esta cantidad de información en constante crecimiento, la cual ya está demasiado grande para poder revisarla durante una vida humana —y mucho menos para la siguiente clase o la fecha de entrega del siguiente artículo—. Sin embargo, parece que ChatGPT no da la talla a lo que necesitamos. Todavía tenemos que actuar con cautela; construir y aplicar nuestro propio criterio crítico; buscar la información original y contrastarla con lo que creemos saber y lo que nos dicen; buscar, comparar y evaluar información proveniente de diferentes fuentes.

Entonces, ¿sueñan las inteligencias artificiales con ser autoras? Primero, un paréntesis explicativo: esta pregunta está inspirada en la novela de ciencia ficción de Philipp K. Dick titulada *Do androids dream of electric sheep?* (Dick, 1968/2020). Describe un mundo distópico donde seres artificiales, originalmente creados para trabajar en mundos extraterrestres de difíciles condiciones y servir a la humanidad, se han infiltrado en la sociedad humana de la tierra y hay el temor de que planean una conspiración. Algunos de los temas son: ¿Qué diferencias hay entre seres inteligentes naturales y seres inteligentes artificiales que justificaran un tratamiento diferente? Si estos seres artificiales tienen habilidades superiores, ¿se volverán una amenaza para la humanidad? Son temas que han (re)surgido en los últimos años, dado que las IA

finalmente parecen cumplir con las expectativas que hubo desde hace 60 años.

Mi respuesta: no, no considero que ChatGPT y otras IA sueñen con ser autoras. Son todavía las personas humanas que sueñan con poder delegar ciertas tareas a estas IA. No es un sueño tonto ni ilegítimo. Solo digo que todavía no estamos allí y que un peligro consiste en creer que sí lo estemos y ya no esforzarnos en usar nuestras capacidades críticas. ChatGPT y otras IA se han podido infiltrar porque corresponden a nuestros sueños, están cambiando las reglas del juego de la producción académica, son disruptivas. Hay que entrar a esta situación con los ojos abiertos y no (solo) soñando, con la disposición de aprender cómo este nuevo mundo se ve, cuáles nuevas opciones sirven y cuáles no. Sí hay oportunidades y sí hay amenazas, pero es nuestra responsabilidad detectar qué es qué.

Finalmente, y cambiando de tema, quisiera recomendarles mirar y apreciar la portada de este número, donde se presenta la primera obra de Zoé Romero Rodríguez, la artista que accedió a acompañarnos en este 2023. "La metáfora del ser humano", como se titula esta obra, a lo mejor nos ayude a reflexionar sobre lo que hace a un ser justamente un ser humano.

Referencias

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, *15*(2), Artículo e35179. https://doi.org/10.7759/cureus.35179

Awan, A. A. (2023, 17 de marzo). *A guide to using chatGPT for data science projects* [Tutorial]. Datacamp. https://www.datacamp.com/tutorial/chatgpt-data-science-projects

Berdejo-Espinola, V., & Amano. T. (2023). AI tools can improve equity in science. *Science*, *379*(6636), 991. https://doi.org/10.1126/science.adg9714

Ballinger, S. (2023). *ChatGPT: Complete chatGPT course for work 2023 (ethically)!* [MOOC]. Udemy. https://www.udemy.com/course/chatgpt-complete-chatgpt-course-for-work-2023-ethically-chat-gpt/

Dick, P. K. (2020). *¿Sueñan los androides con ovejas eléctricas?* Minotauro. (Obra original publicada en 1968)

Elali, F. R., & Rachid, L. N. (2023). AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns*, *4*(3), Artículo 100706. https://doi.org/10.1016/j.patter.2023.100706

Fahrenberg, J. (2011). *Wilhelm Wundt – Pionier der Psychologie und Außenseiter? Leitgedanken der Wissenschaftskonzeption und deren Rezeptionsgeschichte* [Wilhem Wundt – ¿pionero de la psicología y forastero? Ideas rectoras del concepto de ciencia y su historia de recepción]. PsychArchives. https://doi.org/10.23668/psycharchives.10417

Hillier, M. (2023, 20 de febrero). *Why does ChatGPT generate fake references?* Teche. https://teche.mq.edu.au/2023/02/why-does-chatgpt-generate-fake-references/

Lee, J. Y. (2023). Can an artificial intelligence chatbot be the author of a scholarly article? *Science Editing*, *10*(1), 7-12. https://doi.org/10.6087/kcse.292

McAdoo, T. (2023, 7 de abril). *How to cite chatGPT*. APA Style Blog. https://apastyle.apa.org/blog/how-to-cite-chatgpt

Müller, O. (2022). Nada nuevo: la plaga del plagio [Editorial]. *Avances en Psicología Latinoamericana, 40*(2), i-iii. https://revistas.urosario.edu.co/index.php/apl/article/view/12433

Nature. (2023). Tools such as ChatGPT threaten transparent science; here are our ground rules for their use [Editorial]. Nature, *613*, 612. https://doi.org/10.1038/d41586-023-00191-1

Nolan, S. A., & Kimball, M. (2023, 15 de marzo). *Learning to lie: The perils of ChatGPT*. Psychology Today. https://www.psychologytoday.

com/us/blog/misinformation-desk/202303/learning-to-lie-the-perils-of-chatgpt?eml

OpenAI. (2023). ChatGPT (Versión Mar 23) [Modelo de lenguaje grande]. https://chat.openai.com/chat

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. Nature, *613*, 620-621. https://doi.org/10.1038/d41586-023-00107-z

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, *379*(6630), 313. https://doi.org/10.1126/science.adg7879

Wundt, W. (1874). *Grundzüge de physiologischen Psychologie* [Fundamentos de psicología fisiológica]. Wilhelm Engelmann. https://archive.org/details/b21905393/page/n5/mode/1up

Wundt. W. (1900). *Völkerpsychologie: Die Sprache* [Psicología de los pueblos: el lenguaje] (2 Vol.). Wilhelm Engelmann. https://pure.mpg.de/pubman/faces/ViewItemFullPage.jsp?itemId=item_2407695_4

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*, 298-300. https://doi.org/10.1038/485298a

Nota sobre la portada

Título: *Metáfora del ser humano*[1]
Autora: Zoé Romero Rodríguez
Medidas: 120 x 90 cm
Técnica: óleo sobre lienzo
Año: 2020

[1] La autora y exclusiva teniente de todos los derechos sobre esta obra, inclusivo de venta, impresión y reproducción, es Zoé Romero Rodríguez.