La verdadera amenaza de la inteligencia artificial generativa

A pesar de la expectativa que ha generado, la inteligencia artificial generativa (IAG) puede resultar más limitada de lo que se piensa, pero su principal amenaza está relacionada con la extraordinaria capacidad que presenta para generar información falsa y hacerla pasar por verdadera.

> "Lo siento, Dave, pero de acuerdo con la subrutina C1532/4, cito: 'cuando la tripulación esté muerta o incapacitada, el ordenador debe asumir el control', fin de la cita. Debo, por tanto, asumir tu autoridad ahora dado que no estás en condiciones para ejercerla con inteligencia".
> *2001: Una odisea del espacio* (Stanley Kubrick, 1968)

Aunque parece un argumento de película de ciencia ficción, ya desde hace tiempo hay expertos avisando de que la inteligencia artificial se puede volver contra el ser humano. Hace unos años algunas figuras relevantes de la ciencia y la tecnología -como Bill Gates, Elon Musk o el mismísimo Stephen Hawking- alzaron su voz alertando sobre los peligros para la especie que traía consigo el grado de desarrollo actual de esta tecnología.

El último episodio de esta historia ha llegado de la mano de los productos basados en la inteligencia artificial generativa, especialmente aquellos de la empresa estadounidense Open.AI, como el *software* generador de imágenes Dall-e o el chatbox ChatGPT, cuya cuarta versión entró en funcionamiento en 2023 alcanzando un gran impacto social por su capacidad desplegada de generar conocimiento. Como es un sistema abierto, todo el mundo ha podido interactuar con ChatGPT4 y comprobar su habilidad para generar textos sobre cualquier tema sobre el que se le pregunte. De ahí el revuelo mediático.

Durante la primera mitad de 2023, el sistema de Open.AI ha ocupado portadas y titulares, y ha generado una ola de temor a la inteligencia artificial generativa que ha recorrido el mundo. La inteligencia artificial generativa (IAG) es una rama de la inteligencia artificial que crea nuevos resultados -un texto, una imagen, un archivo de sonido- en base a los datos recibidos, a diferencia de los sistemas tradicionales centrados en el reconocimiento de patrones y en la elaboración de predicciones. La primera alarma relacionada con esta tecnología ha sonado en el terreno del empleo: ¿a cuántas profesiones puede desplazar este sistema? ¿qué competencias actualmente desempeñadas por humanos quedarán obsoletas? El segundo motivo de preocupación está relacionado con la maestría con la que los sistemas de IAG pueden crear contenidos falsos que parecen completamente reales. Suponen, por tanto, herramientas eficaces para crear desinformación, *fake news* y *deep fakes*.

Durante la primera mitad de 2023, el sistema de Open.AI ha ocupado

portadas y titulares, y ha generado una ola de temor a la inteligencia artificial generativa que ha recorrido el mundo

Esta vez han sido los propios responsables del desarrollo de este tipo de algoritmos los que han anunciado que traerán consigo el apocalipsis. En una carta publicada el 30 de mayo, 350 ingenieros y ejecutivos de empresas tecnológicas –entre los que se encontraban Sam Altman (presidente ejecutivo de OpenAI), Demis Hassabis (Google DeepMind) y Dario Amodei (Anthropic), las compañías más punteras en este campo- han comparado la amenaza que supone para la humanidad la inteligencia artificial con una guerra nuclear o una pandemia[1]. Para algunos tecnólogos -encabezados por Ray Kurzweil- en algún momento no muy lejano, la inteligencia artificial superará a la humana, y las propias máquinas serán capaces de crear máquinas mucho más inteligentes que las actuales. Es lo que se denomina la *singularidad tecnológica*.

Noticias como la que salto a los medios en junio sobre un dron inteligente militar estadounidense que había decidido matar a su operador en una simulación, porque consideraba que le obstaculizaba para alcanzar los objetivos que tenía fijados, no hacen más que alimentar la histeria sobre las maquinas autónomas y los sistemas inteligentes[2]. Pues bien, varios días después, la propia Fuerza Aérea estadounidense negó públicamente que el incidente hubiera tenido lugar, ni que se hubiese realizado un ejercicio de ese tipo con aeronaves autónomas[3]. Todo ello nos lleva a considerar si el verdadero problema para la humanidad no es la desinformación, y no la supuesta superioridad de la inteligencia artificial.

La pregunta es ¿está realmente tan avanzado el grado de desarrollo de la inteligencia artificial? ¿Podemos hablar de inteligencia real o tan solo estamos ante herramientas estadísticas muy poderosas y sofisticadas?

Los modelos amplios de lenguaje (LLM)

Los sistemas como el popular ChatGPT son lo que se conoce como modelos amplios de lenguaje o *large language models* (LLM). Se trata de herramientas de inteligencia artificial basadas en lógicas bayesianas que identifican pautas de lenguaje en la información que reciben, ya sea texto u otras, y devuelve la respuesta en forma de textos bastante bien escritos. Como indica Andrés Ortega, investigador del Real Instituto Elcano, los LLMs dan la impresión de aprender y de usar representaciones del mundo, pero realmente se trata en gran medida de mera imitación o copia directa de textos u otros productos realizados por seres humanos. Y el problema adicional es que con frecuencia se nutren de fuentes poco fiables, cuya validez no son capaces de determinar, por lo que los resultados que arrojan de las consultas deben ser tratados con cautela.

Los sistemas como el popular ChatGPT son lo que se conoce como modelos

amplios de lenguaje o large language models (LLM)

Sobre esto último, el gurú de la IA Gary Marcus avisa de los peligros que implica utilizar Chat GPT: “primero, no es muy fiable, para una misma pregunta, a veces da información correcta, otras veces no. Segundo, tiene el problema que en IA llamamos alucinaciones, se inventa información y no hay ninguna señal que avise de que se está inventando algo”[4]. Marcus está convencido de que este chatbot de Open.IA tendrá un impacto social, si bien todavía no tiene claras sus posibles aplicaciones, aparte de “escribir trabajos escolares”. Lo que le preocupa es el potencial que presenta para producir desinformación a través de la creación de contenidos falsos.

El filósofo Slavoj Žižek va todavía más lejos que Marcus cuando proclama que este tipo de chatbots a menudo son infantiles y estúpidos, pero no lo suficientemente infantiles y estúpidos como para pillar los matices, la ironía y las contradicciones inherentes a la cultura y la comunicación humanas. A su juicio, el interactuar con este tipo de sistemas nos puede convertir en igual de obtusos que ellos. A modo de ejemplo, Žižek se pregunta si un chatbot entendería que la frase “compre una cerveza por el precio de dos y llévese la segunda gratis” es un chiste, una ironía evidente.

Las limitaciones del aprendizaje automático

Gary Marcus considera que los miedos y recelos sobre la inteligencia de sistemas como ChatGPT son producto de la sobre-atribución, es decir, el atribuir una vida mental a estas máquinas que no existe en realidad. Cuanto mejor se comunican los LLMs en nuestro lenguaje más tendemos a atribuirles cualidades humanas, por ejemplo, utilizando para referirnos a ellos términos como “sabe”, “reconoce” o “pìensa”, o, peor aún, atribuirles conciencia. Murray Shanahan, del Imperial College de Londres, afirma que estos modelos de lenguaje lo único -que no es poco- que hacen es generar secuencias de palabras estadísticamente afines o más probables en función de las consultas recibidas, con la información y los textos de que disponen, o que han sido utilizados para su aprendizaje. Por eso, resulta ridículo compararlos con la mente de una persona repleta de intereses, esperanzas y deseos, aunque estén diseñados para interactuar con nosotros como si fueran de los nuestros.

El matemático Neil Saunders recuerda que Alan Turing, uno de los padres de la informática, decía que un ordenador no necesita comprender un algoritmo para ejecutarlo, y, de igual forma, el hecho que ChatGPT sea capaz de escribir en un lenguaje pasional no significa que entienda el significado de las frases que genera. Es un fenómeno que se ha definido muy acertadamente como “competencia sin comprensión”. Al igual que Shanaham, Saunders avisa del peligro de otorgar rasgos antropomórficos a los algoritmos: “debemos darnos cuenta de que no son más que máquinas probabilísticas sin intenciones o preocupación por los humanos”. A pesar de las mayores capacidades que presenta la versión 4 de ChatGPT, no podemos habar de que manifieste ni comprensión, ni intenciones, nada más que “aplicación de patrones”.

Alan Turing, uno de los padres de la

informática, decía que un ordenador no necesita comprender un algoritmo para ejecutarlo

El experto Judea Pearl en sus estudios sobre inteligencia artificial establece tres niveles de la que denomina la *escalera de la causalidad*, que acercaría el funcionamiento de las máquinas al razonamiento humano. A su juicio, los algoritmos de *deep learning* actuales se quedan en el primer peldaño, el de aprendizaje por asociación: acumulan grandes cantidades de datos que les permiten establecer la probabilidad de que exista una correlación entre dos cosas. Sin embargo, no son capaces de determinar la relación causa y efecto entre ellas, si existe. La verdadera inteligencia llegará cuando estos sistemas sean capaces de especular sobre acciones que no tienen un precedente, y así poder responder a preguntas más complejas -la fase de intervención-, y, finalmente, cuando puedan imaginar situaciones alternativas, es decir, contestar a preguntas como ¿habría ocurrido un evento si otro evento en el pasado no hubiera tenido lugar? Es la fase retrospectiva, porque se basa en imaginar el presente con un pasado distinto. Así que aún le queda mucho camino por recorrer a la máquina inteligente.

El verdadero peligro de la inteligencia artificial generativa

A pesar de su indudable capacidad para crear contenidos, no parece que los modelos actuales de inteligencia artificial generativa vayan a desplazar por ahora a la inteligencia humana. Es verdad que determinadas habilidades que presentan pueden suponer una amenaza para algunas profesiones, pero por automatizables que resulten algunas tareas siempre deben llevar supervisión humana. Los LLMs podrán redactar textos legales o periodísticos mucho más rápido que los abogados y los redactores, y, sin embargo, no tienen la capacidad para evaluar la validez de su producto, que debería ser revisado por un experto en el tema.

La supuesta creatividad de los modelos amplios de lenguaje también es algo más que cuestionable, dado que lo que suelen hacer es fundir en sus resultados a las consultas recibidas contenidos de distintas fuentes. En este sentido, Charlie Brooker, el creador de la popular serie de Netflix *Black Mirror*, realizó el curioso experimento de pedirle a ChatGPT que escribiese el guion de un nuevo capítulo. El resultado no pudo ser más decepcionante: el algoritmo había cogido las sinopsis de distintos capítulos y las había mezclado para escribir su trama[5].

El verdadero peligro de este tipo de inteligencia artificial es su utilización para crear bulos y mentiras para inundar los espacios públicos de desinformación que pueda manipular la opinión pública. El mal uso de la inteligencia artificial generativa por los propios humanos es la mayor amenaza que presenta en este momento para nosotros. Este es el aspecto que necesita de mayor legislación y control institucional. Incluso se ha llegado a sugerir que sería conveniente que cualquier contenido generado por inteligencia artificial lleve una marca de agua que indique al interlocutor que se está tratando con un chatbot.

El verdadero peligro de este tipo de

inteligencia artificial es su utilización para crear bulos y mentiras para inundar los espacios públicos de desinformación que pueda manipular la opinión pública

La desinformación y el engaño que proliferan por los medios de comunicación y, especialmente, a través de las redes sociales supone una seria amenaza para la democracia. Michael Sandel, entrevistado por la revista Telos, advierte del riesgo que supone que las generaciones que vienen pierdan interés por la línea que separa lo que es falso de lo que es real. En sus palabras: “este es el verdadero peligro, no solo que cada vez sea más difícil distinguir lo que es real de lo que es falso, sino que esa distinción deje de importarnos”. Controlar la utilización que se hace de los algoritmos generativos resulta crucial en este sentido.

Foto de Rahul Pandit

Fundación Telefónica (2023) Revista Telos nº 122. *Posverdad***.** Junio 2023. Disponible en: https://www.fundaciontelefonica.com/cultura-digital/publicaciones/telos-122-posverdad/785/

Hackl, C. (2023) “¿Qué es la inteligencia artificial generativa y qué significa para tu marca?” en *Wired.* Disponible en: https://es.wired.com/articulos/que-es-la-inteligencia-artificial-generativa-y-que-significa-para-las-marcas

Luccioni, S. y Marcus, G. (2023) “Stop Treating AI Models Like People” en *The Road to AI We Can Trust.* Disponible en: https://garymarcus.substack.com/p/stop-treating-ai-models-like-people

Ortega, A. (2023) “Nuevas inteligencias” en *Telos.* Disponible en: https://telos.fundaciontelefonica.com/nuevas-inteligencias/

Pearl, J. (2018) “The Book of Why: The New Science of Cause and Effect”.

Saunders, N. (2023) “Evolution is making us treat AI like a human, and we need to kick the habit” en *The Conversation*. Disponible en: https://theconversation.com/evolution-is-making-us-treat-ai-like-a-human-and-we-need-to-kick-the-habit-205010

Shanham, M. (2022) “Talking About Large Language Models”. Disponible en: https://arxiv.org/pdf/2212.03551.pdf

Žižek, S. (2023) “Artificial Idiocy” en *Project Syndicate*. Disponible en: https://www.project-syndicate.org/commentary/ai-chatbots-naive-idiots-no-sense-of-irony-by-slavoj-zizek-2023-03