Verdad y endogamia en las inteligencias artificiales generativas

Por qué una IA nunca creará un Nietzsche

MIGUEL PALOMO

La preocupación por la convivencia de la sociedad presente y futura con máquinas avanzadas como las inteligencias artificiales generativas (IAG) se palpa en el ambiente. De hecho, es habitual que en estos días encontremos todo tipo de conjeturas sobre cómo las IAG van a cambiar nuestro mundo. Para unos, nos encontramos con un apocalipsis social; para otros, el futuro próximo será una utopía en la que apenas realizaremos trabajo de gestión. Valga entre ambos extremos, por supuesto, todo tipo de posiciones. Sin embargo, viendo que la convivencia con las IAG es una realidad, creo que sería más fructífero realizar un análisis razonable sobre lo que por definición podemos y no podemos esperar de las IAG independientemente de su desarrollo futuro, así como el tipo de relación que tenemos con ellas y qué problemas puede conllevar todo ello. Verdad, realidad, confianza, desinformación y endogamia son términos que se nos antojan necesarios para comprender esta compleja y cada vez más estrecha relación que mantenemos con las máquinas.

Veamos primeramente respondiendo qué podemos esperar de las IAG con un ejemplo sencillo: si preguntamos a ChatGPT (la IAG más conocida y utilizada a día de hoy) sobre la inauguración el Canal de Panamá en 1914, nos ofrecerá una serie de datos que pretenden explicar qué ocurrió en aquel año. Damos por hecho que la información es *correcta*, lo que viene a significar que las afirmaciones ofrecidas por la IAG son *verdaderas* en tanto que se corresponden con la *realidad*. Si nos habla de Panamá, pensaremos en el país de nombre idéntico. Si nombra el número 1914, pensaremos en ese año dentro de la línea cronológica de la historia. Quizá haya quienes duden de la información que aparece en pantalla y busquen comprobar los datos a través de otras herramientas digitales. Sin embargo, la cuestión es algo más compleja, puesto que la veracidad de la información que nos

proponen las IAG es relativa y estas ni siquiera tienen capacidad de referirse a la realidad como tal.

Comencemos por la cuestión de la veracidad de la información que nos ofrecen las IAG. En su estado actual, ChatGPT, Bing y otras IAG se inventan referencias, personajes, fechas e incluso títulos de libros que nunca han existido. Es habitual defender que con el continuo desarrollo de las IAG este problema quedará solucionado en breve. Sin embargo, puesto que su fuente es la información existente en internet (incluyendo por un lado libros y artículos académicos, pero por otro lado también foros de internet o redes sociales), ¿qué garantía tenemos de que la selección realizada por el algoritmo filtre adecuadamente aquello que no es veraz? En realidad, lo único que nos queda es la *confianza* hacia los responsables del algoritmo.

A su vez, podemos pensar en varias objeciones que retan esta percepción de que el desarrollo de las IAG solventará este problema. Por un lado, no resulta alocado imaginar que se pueda utilizar el funcionamiento del algoritmo de ChatGPT para que ofrezca cierto tipo de información en lugar de otra, sin siquiera acceder al algoritmo como tal ni conocer estrictamente su funcionamiento exacto. Ya hemos tenido casos similares como, por ejemplo, conseguir que el buscador de Google te ofrezca unos resultados u otros al hacer una búsqueda. Sería tan sencillo como descubrir que ante unos *inputs* la IAG ofrece unos *outputs* concretos, y ofrecer información en diversos lugares de la sociedad digital para conseguir que el barrido de información realizado por *bots* acoja esta información como veraz. Por lo tanto, el desarrollo como tal o la mejora de la herramienta no es suficiente para solucionar este problema. La desinformación tiene su éxito precisamente en su pretensión de veracidad, y las IAG no son ajenas a la manipulación y a la persuasión externa.[1]

La desinformación tiene su éxito precisamente en su pretensión de veracidad, y las IAG no son ajenas a la manipulación y a la persuasión externa

El segundo problema tiene un corte filosófico. Es evidente que las IAG no pueden hacer referencia a la *realidad*: solo ofrecen símbolos y palabras sin significado para el algoritmo, aunque a nosotros nos parezcan frases perfectamente comprensibles. Si preguntamos sobre el Canal de Panamá y la IAG nos responde que

[1] Miguel Palomo, «Incidencias filosóficas actuales en la sociedad digital: ideologías, desinformación y confusión epistemológica», *Arbor*, 197(802), 2021, a630. https://doi.org/10.3989/arbor.2021.802008

se inauguró en 1914, ¿está haciendo la IA referencia a un hecho histórico que tuvo lugar en un momento específico de la historia? Solamente hay que recuperar algunos ejemplos clásicos de la filosofía para poner esto en duda. Tal es el caso de la propuesta de Hilary Putnam[2] que nos hablaba de una hormiga que, paseando por la arena azarosamente, dibuja sin percatarse el rostro de Winston Churchill. ¿Está haciendo la hormiga referencia a Churchill? Habría que responder que no. Una serie de líneas no representan nada por sí mismas, del mismo modo que un carácter por sí mismo tampoco hace referencia a nada: se trata de líneas contingentes. Similar es el ejemplo de un conjunto de simios golpeando una máquina de escribir: aunque azarosamente lograsen escribir un discurso perfecto, estas palabras no harían referencia a nada. Del mismo modo ocurre actualmente con las IAG. El contexto que rodea a la IAG nos persuade haciéndonos pensar que hay intención y representación, cuando en realidad ambas están ausentes. El hecho de que humanicen a ChatGPT por ejemplo, haciendo como que escribe en un teclado al estilo humano; que hable educadamente; o que se disculpe cuando no tenga una respuesta; todo ello enmascara que la IAG, simplemente, ante un *input* (nuestra pregunta) está entrenada para ofrecer un *output*: un conjunto de símbolos a los que otorgamos la etiqueta de "respuesta", pero que para la máquina no tiene sentido alguno. Por lo tanto, nos encontramos con dos importantes problemas: las IAG no ofrecen ni veracidad ni conexión con la realidad.

Siendo así, parece legítimo preguntarse, ¿qué relación debemos tener con las IAG? O dicho de otro modo, ¿por qué confiar en estas máquinas de *software*, si no ofrecen verdad ni veracidad? ¿Por qué las utilizamos? Es más, si los algoritmos son manipulables internamente, de modo que los creadores del algoritmo escojan específicamente qué información ofrecer ante algún tipo de propuesta polémica; y si también son manipulables externamente, de modo que consigamos colocar la información que le interese a algún grupo de agentes en los *outputs* de las IAG, ¿qué razón tenemos para que siga existiendo un lazo de confianza en la información que nos proveen estas herramientas digitales? Sin pretensión de que suene peyorativo, aunque sí realista, la razón que tenemos para confiar en las IAG es la tendencia humana hacia la ingenuidad; nuestra tendencia hacia la fe en discursos que presentan un patrón o una pretensión de veracidad. Dicho de un modo más riguroso, la necesidad de descansar en una epistemología del testimonio, es decir, de confiar en el testimonio de terceras personas (o tecnopersonas, en el caso de las IAG). El ser humano necesita esta epistemología para algo tan importante

[2] Hilary Putnam, *Razón, verdad e historia*, Tecnos, Madrid, 1988.

como el avance en el conocimiento. Un científico, por ejemplo, necesariamente confía en el trabajo realizado por otros; un historiador confía en fuentes fiables que hablen sobre hechos pasados; un lector confía en la intencionalidad del escritor. Estos patrones de confianza son necesarios para la vida práctica y aparecen en discursos ideológicos, sociales, filosóficos o religiosos, entre otros muchos. Puesto que parece ser una forma en la que los humanos funcionamos, esta confianza se vuelca ahora, para bien o para mal, en las IAG.

¿Qué es la información endogámica?

La confianza, sin embargo, no debe ser ciega. Por definición propia de lo que son las IAG (y que seguirán siendo por mucho desarrollo y mejora que se les aplique) surge otro problema, que es lo que denomino *información endogámica*. Las IAG gestionan cantidades ingentes de información. La clave creo que está en el verbo que he utilizado: gestionar, es decir, tomar información ya existente y darle una forma u otra, según el *input* recibido. Esto quiere decir que una IAG no genera información nueva, sino que da forma a la información recogida; es decir, moldea una información ya procesada (el conjunto de información barrida por *bots*) para ofrecer la información solicitada en el *input*. Es cierto que las IAG pueden gestionar tal cantidad de información que incluso pueden proponer predicciones probabilísticas, lo cual en ocasiones ha resultado en un avance en diversos tipos de investigaciones. Encontramos un ejemplo en una investigación reciente que sugiere que una IA permite detectar un 20% más de positivos de cáncer de mama que el método habitual de doble confirmación por parte de radiólogos.[3] Sin embargo, aunque importante, ese ejercicio solamente muestra una gestión de la información recibida mediante algoritmos que posteriormente requiere de la confirmación de un experto médico para confirmar o desmentir la existencia de un cáncer de mama. Por lo tanto, la IAG gestiona información, y en base a esta, realiza predicciones.

Nos encontramos con dos importantes problemas: las IAG no ofrecen ni veracidad ni conexión con la realidad

Pero cabría preguntarse, ¿puede ofrecer información nueva? Me veo en la obligación de decir que no. Por mucho que avance el desarrollo de las IAG, estas no

[3] Kristina Lång, Viktoria Josefsson, Anna-Maria Larsson, Stefan Larsson, Charlotte Högberg, Hanna Sartor, Solveig Hofvind, Ingvar Andersson, Aldana Rosso, «Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study», *The Lancet Oncology*, vol. 24, núm. 8, 2023, pp. 936-944, https://doi.org/10.1016/S1470-2045(23)00298-X.

pueden salirse del *box* (o conjunto de contenidos) de información que se les ha ofrecido. Por ejemplo: la IAG podrá ofrecer información sobre Newton, Einstein, Born o Schrödinger; podrá utilizar datos de millones de artículos científicos para ayudar a buscar la teoría que unifique la teoría de la relatividad general de Einstein con la mecánica cuántica; podrá explicarnos la filosofía de Platón para que la entienda un niño de 9 años o quizá un especialista en pensamiento griego. Sin embargo, una IAG no puede ofrecer un patrón de interpretación sobre los paradigmas científicos o filosóficos y proponer uno nuevo. El nivel de abstracción necesario para realizar algo así requiere que el sujeto protagonista de este ejercicio esté por encima del *box* de información, y por supuesto que añada algo que no estaba en él. La IAG, en su gestión, lo único novedoso que puede aportar es resultado de información ya presente, pero que debido a nuestras limitaciones de cálculo no hemos podido llegar a identificar (podríamos pensar en el cálculo de los decimales del número pi, por ejemplo). Entonces, ¿qué puede hacer una IAG? Regurgitar información, reciclarla, traducirla, gestionarla. Este reciclaje continuo es útil, pero limitado. Esta limitación, natural por otro lado, no puede dejar de poseer un matiz sospechoso para una mente crítica. En tanto que las IAG solo pueden hacer referencia a la información recogida, esto fomenta una endogamia en la información, de modo que lo que se ofrece es resultado de la información ya disponible para las IAG. Podemos pensar en el siguiente ejemplo: pedir a ChatGPT que nos resuma las *Meditaciones metafísicas* de Descartes[4] y utilizar esa información, digamos, reciclada, para tareas prácticas, como redactar un manual de historia de la filosofía que previsiblemente será subido al entorno digital y por lo tanto es susceptible de ser barrido por *bots* y procesado por las IAG. No es mi intención realizar una valoración moral, es decir, señalar si esto es deseable o no. Las bondades y los miedos sobre lo que puede ofrecer o no las IAG la gran mayoría de veces resultan ser ejercicios de *marketing*. Lo que acaece, los hechos, son moralmente neutros. Y en este sentido la información endogámica es un hecho.

Cabría preguntarse, ¿acaso no es lo mismo reciclar información a través de las IAG que recogerla de libros tradicionales, periódicos en papel o incluso de las antiguas enciclopedias? Podemos pensar en el caso del escritor de un manual sobre historia de la filosofía, pero en esta ocasión habiéndose informado solamente por medios tradicionales. Se entiende que este debe acudir a multitud de obras ya escritas por diferentes pensadores a lo largo de la historia, simplificar y moldear esa información ya presente en otros lugares, de modo que resulte com-

[4] René Descartes, *Discurso del método; Meditaciones metafísicas*, Espasa-Calpe, Madrid, [1637] 2004.

prensible para un público con poca o nula especialización. Sin embargo, habría que responder que el caso sí difiere con respecto al de las IAG, porque en la automatización de las máquinas se excluye la racionalización humana. Los algoritmos "eligen" basándose en (a) información previa; (b) las indicaciones propuestas por los creadores del *software*; (c) el *input* del sujeto que solicita cierta información. En la selección del *output* adecuado, la máquina no puede realizar un ejercicio de racionalización, en tanto que la racionalización humana es capaz, primeramente, de saltarse los pasos que se le indiquen, y además, aportar algo nuevo que no estaba presente ni en (a), ni en (b) ni en (c). Supongamos el mismo caso del autor que se encuentra escribiendo un manual de filosofía. Tras la lectura y reflexión sobre lo leído, quizá encuentre una clave que ha resultado invisible para otros (tal y como suele ser el espíritu del tiempo de cada época) y que resulte esencial para trazar una línea que explique por qué unas formas de pensar y comprender el mundo han seguido históricamente a otras. Nos encontramos en este caso con una mirada novedosa, con información nueva que no resulta endogámica. Podríamos hacer una analogía con un conocido caso literario en la ciencia ficción. Los simios de *2001: una odisea en el espacio*, una vez evolucionan, cuando ven un hueso, consiguen darle un significado distinto a su sentido original. Este ya no es hueso: se trata de una herramienta que puede servir para matar. Ese cambio requiere un tipo de salto facilitado por un proceso de racionalización. Análogamente, si volvemos al caso de las IAG, al carecer de racionalización, no son capaces de salir de aquello que se les ha predeterminado por diseño previo. Es por ello que el escritor del manual de historia de la filosofía podrá presentar una nueva perspectiva basada en su comprensión y racionalización de lo que está leyendo en libros escritos previamente, mientras que la IAG simplemente gestionará la información recogida.

Las IAG solo pueden hacer referencia a la información recogida; esto fomenta una endogamia en la información

Podría pensarse que los últimos avances propuestos por las empresas que han creado IAG podrán solucionar esta información endogámica. Por ejemplo, OpenAI muy recientemente ha ofrecido la posibilidad de crear "ChatGPTs", es decir, escisiones de la IAG que son capaces de especializarse en información concreta que un sujeto les aporte.[5] Por ejemplo, puedo entrenar a un ChatGPT con información

[5] Alex Heath, «OpenAI is letting anyone create their own version of ChatGPT», *The Verge*, 6 de noviembre de 2023, disponible en: https://www.theverge.com/2023/11/6/23948957/openai-chatgpt-gpt-custom-developer-platform

de publicaciones académicas sobre filosofía para convertirse en una IAG especializada en ofrecer información filosófica, siguiendo órdenes específicas realizadas por mí; puedo entrenarlo para que me ofrezca las mediciones correctas en planos de arquitectura; o quizá entrenarlo en la búsqueda de patrones propios de IAG en la corrección de trabajos universitarios redactados por discentes. Podría decirse que, siguiendo la información nueva que yo puedo aportar, las IAG ya no están limitadas a un *box* cerrado de información. Pero, a pesar de ello, las IAG siguen dependiendo de la información que haya en ese *box*, aunque sea un sujeto el que aporte información nueva. De hecho, esto viene a confirmar que toda información nueva proviene de sujetos externos a la IA. Pero, ¿y si esa información que volcamos viene determinada de algún modo por la información que las IAG nos ofrecen? Ello nos lleva de vuelta a la pregunta sobre cómo debe ser la relación entre humanos e IAG.

El pensamiento endogámico

En un intento por adelantarnos al futuro próximo, creo que nuestra relación con estas máquinas debe ser de prudente cautela, debido a que es muy fácil que la información endogámica acabe ocasionando un pensamiento endogámico: un reciclaje continuo de lo ya pensado y propuesto en el pasado. ¿Cómo se pone en práctica el pensamiento endogámico? Habría tres pasos: la alimentación, la retroalimentación y la persuasión ideológica. El primer paso se produce mediante los ejercicios que permiten la acumulación de información por parte de las IAG, que ya he comentado. El segundo se produce en el proceso de *inputs* y *outputs* por parte de los usuarios, que también podemos denominar simplemente proceso de preguntas y respuestas. En este momento, el sujeto interactúa con la IAG de modo que consiga la información que originalmente estaba buscando. De este modo, las respuestas aparecen como un *output*: se trata, básicamente, de información endogámica, y se produce por una retroalimentación continua entre las IAG y sus usuarios. El último paso es el más relevante y el que más interés tiene para la filosofía: que la información endogámica pase a crear pensamiento endogámico. Puesto que la información es endogámica, si el pensamiento se vale de dicha información, es muy probable que acabemos con un pensamiento que difícilmente sea capaz de trascender esa endogamia. Al igual que la tesis de Nicholas Carr, quien se preguntaba si Google nos hacía más estúpidos,[6] quizá las IAG nos hacen,

[6] Nicholas Carr, «Is google making us stupid?», *The Atlantic Monthly*, 302(1), 2008, pp. 56–58.

como mínimo, menos propensos a la abstracción y a la pausa que requiere un razonamiento adecuado para enfrentarse a problemas de capital importancia.

Es precisamente tarea de la filosofía, no el comentar hasta la saciedad lo ya dicho en el pasado por tal o cual filósofo, sino ser capaz de pensar nuestro tiempo. Lo primero es resultado de una información endogámica; lo segundo, de la puesta en práctica del razonamiento humano, que no se alcanza cuantitativamente (pudiendo almacenar y gestionar cantidades ingentes de información), sino cualitativamente (pudiendo incluso romper las normas de lo establecido previamente). La filosofía de nuestro tiempo, precisamente, debe superar las limitaciones técnicas que imponen otros modos de pensar: tal es el caso de las IAG.

Nuestra relación con estas máquinas debe ser de prudente cautela, debido a que es muy fácil que la información endogámica acabe ocasionando un pensamiento endogámico

El filósofo Carlos Pereda hablaba del sucursalismo como uno de los problemas de la filosofía en la actualidad: el reiterar ciertas fórmulas repetidas una y otra vez "hasta vaciarlas por completo de contenido", formas de pensamiento fruto de la "pereza teórica" y a cuya producción llama "cultura de secta".[7] A esto se le añade, según Pereda, un temor a cualquier cambio, lo que se traduce en una coacción sobre las formas que se consideran correctas de proceder en el pensamiento, o dicho de otro modo, una forma de opresión intelectual. En nuestro contexto también podemos hablar de sutiles formas de opresión en el intento de dominar el pensamiento, o como Javier Echeverría y Lola S. Almendros han llamado: la dominación de las mentes.[8] No es difícil comprender el modo en el que la información volcada y seleccionada en *outputs* va determinando poco a poco las formas de aprehensión de la realidad. En ese sentido, las IAG nos ofrecen un pensamiento, llamémosle, "cómun", resultado de un reciclaje continuo de información, mientras que el pensamiento de nuestra época debe trascender las posibilidades que las IAG nos brindan: superar el espíritu de nuestro tiempo es la única forma de no anclarse a él.

Podemos resumir todo lo dicho de un modo ilustrativo: una IAG nunca podrá crear un Nietzsche, una forma de comprender la realidad que sea capaz de romper los

[7] Carlos Pereda, «La filosofía en México en el siglo XX: un breve informe», *Theoría. Revista Del Colegio De Filosofía*, (19), 2009, pp. 89-108, disponible en: https://doi.org/10.22201/ffyl.16656415p.2009.19.978

[8] Javier Echeverría y Lola Almendros, *Tecnopersonas: como las tecnologías nos transforman*, Ediciones Trea, Gijón, 2020.

moldes de lo previamente propuesto. Para ello se requiere algo más que una información endogámica: hay que tener información, aprehenderla, situarla críticamente en un contexto temporal con respecto a lo que otros han dicho, razonar, intuir e innovar. La información endogámica es producto de una gestión, lo cual no es suficiente para ello. Si la sociedad careciese de personas capaces de trascender el pensamiento de su época, no podría haber saltos entre formas generacionales de pensamiento, lo cual conllevaría un estancamiento filosófico y social. No olvidemos que los grandes descubrimientos científicos y filosóficos surgen, precisamente, en estos saltos generacionales o de épocas. La IAG, por definición, no puede crear patrones de interpretación de la realidad que son los que hacen que pasemos de un paradigma interpretativo a otro, lo que tradicionalmente se ha denominado el paso de una era a otra: de la Edad Media a la Modernidad, a la Ilustración, al mundo contemporáneo, etc.

Conclusiones

Entonces, ¿qué cabe esperar de las IAG? Nada más que una gestión formal de contenidos aportados previamente por sujetos humanos (cabría añadir: y nada menos). Flota en el ambiente la sensación de que todo ha cambiado desde la aparición de ChatGPT: el profesorado se apura en cambiar sus metodologías; los y las trabajadoras temen por sus puestos; la justicia comienza a recibir demandas por diversa información personal (por ejemplo, en forma de imágenes) volcada en las IAG con motivos cuestionables. Y si todo ha cambiado tanto en solo unos pocos meses, ¿qué nos espera en los siguientes?

A ello hay que sumar los problemas que pueden preverse a medio y largo plazo, como son la información y el pensamiento endogámico, sin obviar importantes cuestiones de índole laboral, social y política. Surge de nuevo, por tanto, la pregunta: ¿cómo debemos relacionarnos con las IAG? He denominado la relación como prudente cautela, que viene a significar que la sociedad digital debe ser consciente de la posibilidad real de que diversas olas de desinformación modifiquen los *outputs* y determinen ideologías, cosmovisiones y realidades en la sociedad urbana. El pensamiento crítico, por tanto, se antoja más necesario que nunca.

Miguel Palomo García es profesor de filosofía en la Universidad Complutense de Madrid.