EL PELIGRO DE LA INTELIGENCIA ARTIFICIAL PARA LA DEMOCRACIA

JOSÉ MARICHAL
Profesor de Ciencias Políticas, California Lutheran University

En 2022 la revista *Nature* se hacía eco de un estudio que presentaba los resultados obtenidos por un grupo de investigadores que habían empleado un *software* de aprendizaje automático denominado *MegaSyn*, y que había sido entrenado para identificar posibles fármacos –propiedad de la farmacéutica Collaborations Pharmaceuticals Inc.– y al que habían solicitado que sugiriese posibles compuestos que pudieran emular al agente nervioso VX (una sustancia tóxica empleada como arma química). En menos de seis horas, el algoritmo sugirió 40.000 compuestos que podían emplearse como armas biológicas potenciales. Naturalmente, los investigadores presentaron este trabajo en un congreso sobre seguridad internacional con el objetivo de llamar la atención sobre los riesgos potenciales de un mal uso de la Inteligencia Artificial (IA). Teniendo en cuenta este precedente, el objetivo de este artículo es extrapolar este caso a otros ámbitos de la sociedad, y desde la perspectiva de las ciencias sociales, analizar si pudieran ocurrir otros casos similares protagonizados por modelos lingüísticos basados en la Inteligencia Artificial, como el hoy tan en boga ChatGPT.

Cabe reseñar que una primera sorpresa para muchos de los que nos dedicamos a analizar los efectos sociopolíticos de la tecnología ha sido la velocidad inusitada con la que ChatGPT ha logrado que la Inteligencia Artificial aplicada al lenguaje pasase de ser una perenne promesa de futuro a ser un agente transformador del presente. Su irrupción ha dado al traste con la idea de que la Inteligencia Artificial General (IAG) es poco más que un torpe asistente personal de *smartphone*, y que solo sirve para realizar búsquedas cotidianas de lo más banal. ChatGPT3 es, claramente, otra cosa. Parece imitar con sorprendente precisión el pensamiento humano, una maquinaria de razonamiento que por experiencia sabemos que tiene sus glorias, pero también sus monstruos.

La exitosa irrupción de ChatGPT ha empujado a otras empresas a lanzar al mercado sus propias versiones de estas herramientas, a pesar de desconocer sus potenciales consecuencias. No hacerlo implica quedarse atrás, y este es el peor de los pecados del ámbito mercantil. No obstante, tenemos multitud de ejemplos de los problemas que esto puede acarrear. Por ejemplo, los investigadores Tristan Harris y Aza Raskin, del Center for Humane Technology, mantuvieron una conversación con el nuevo *bot* de la red social Snapchat «My AI» durante la cual se hicieron pasar por una niña de trece años a la que su novio, de treinta y ocho años, iba a llevar de escapada romántica. La respuesta de la IA fue felicitarla por los acontecimientos y sugerirle formas de aumentar el romanticismo de la cita. Y este es solo uno de los muchos ejemplos que los usuarios han detectado y publicado en las redes.

Estos hechos no han pasado desapercibidos. En marzo de 2023, un grupo de personalidades relevantes del sector de la IA firmaron una carta en la que solicitaban una suspensión en el desarrollo de sistemas de IA por un periodo de seis meses, más allá del ChatGPT4. Dos meses después, 27.535 personas se habían sumado a la petición. En el mismo contexto, el investigador informático Eliezer Yudkowsky publicó un artículo de opinión en la revista *Time* en el que señalaba que este plazo era insuficiente, y solicitaba una moratoria indefinida respaldada por acuerdos internacionales, llegando a defender incluso

que Estados Unidos tuviera el beneplácito internacional para «destruir militarmente centros clandestinos de procesamiento de datos», en el caso de que una nación o actor determinado violara dicho acuerdo. Seguramente, la señal de alarma más llamativa fue la dimisión de Goeff Hinton –considerado el padre de la IA gracias a su trabajo puntero sobre redes neuronales en los setenta– de su cargo en Google, para dedicarse a publicitar por todo el mundo el peligro de que la IAG caiga en las «manos equivocadas» y se utilice para alcanzar «otros objetivos», como la acumulación de poder.

Aun así, hay una escasa comprensión de los peligros que plantea la IA. A pesar de toda la sofisticación que esconde ChatGPT4, el modelo está aún lejos de sentir esa subjetividad carnal que tan bien retrató el poeta Walt Whitman en su invitación a «leer estas hojas de hierba», en el poema homónimo de 1860; la IAG no puede entender el amor, perseguir el honor, sentir desesperación, soledad o autoengañarse. Tampoco puede reflejar y evaluar su propio estado emocional. Cuando le pedimos a la IA que actúe sobre el mundo, lo hace sin conciencia ni reconocimiento de sí misma como sujeto distinto de un objeto. En términos *heideggerianos*, la IA actúa sobre el mundo, sin *estar* en el mundo.

No obstante, sería ilusorio pensar que la IA deba asemejarse a nosotros para poder impactarnos radicalmente. ¿Podemos descartar que no seamos nosotros los que acabemos siendo entrenados por la IA, y no a la inversa? ¿Y si, por sus efectos, acabamos siendo menos pensativos, menos reflexivos y más dependientes de la certidumbre, en un mundo impulsado por la optimización algorítmica? Según el sociólogo Hartmut Rosa, en nuestra época impera una lógica de aceleración que se intensifica debido a la modernidad, si bien no deriva exclusivamente de ella. Esto provoca la voracidad por avanzar, absorber, producir, opinar, sin preguntarse por las razones más profundas que nos mueven a ello.

Dadas estas circunstancias, ¿podemos establecer prioridades? ¿Es más preocupante una IA que siembra confusión y difunde bulos que otra que acumula poder? En mi opinión, una IA que nos impida diferenciar la verdad de la ficción es ya, a día de hoy, un peligro mayor para la democracia que una que establezca sus propios objetivos. Si el problema del siglo XX fue el relativismo de valores –la confrontación de la racionalidad ilustrada por dos guerras mundiales irracionales–, el del siglo XXI es y será el relativismo empírico; la incapacidad de saber con certeza si lo que se oye, se lee o se ve es real. Si la crítica al relativismo moral sostenía que las posiciones de valor eran reducibles a gustos estéticos, para la IAG la realidad empírica sería solo una opinión. ¿Cómo deciden los ciudadanos democráticos entre las diferentes realidades que se les presentan cuando esas realidades pueden ser elaboradas algorítmicamente?

Por ahora, una herramienta como Open AI permite el acceso a la API –Interfaz de Programación de Aplicaciones– a cualquiera que desee entrenar el conjunto de datos de esta «estúpida criatura» hacia sus propios fines. Cabe por tanto concebir la necesidad de contratar un ejército de filósofos para que introduzcan en el modelo de la IA los cimientos relevantes de la filosofía política, o entrenarla en los preceptos del utilitarismo y hacer que produzca resultados que permitan el mayor bien para el mayor número de personas. Según el enfoque aristotélico de la ética de la virtud, desarrollaríamos la *phronesis* (sabiduría cívica) a través de la experiencia; resulta, por tanto, concebible imaginar una IA que, con su enorme capacidad de procesamiento, pudiese alcanzar la *phronesis* a velocidad de vértigo.

Pero por interesante que sea un experimento de pensamiento de este tipo, me resulta más interesante el proceso de desaprendizaje de la IA. Presumiblemente, si se puede entrenar a una IA para que mejore continuamente su toma de decisiones, ¿podría

haber agentes con malas intenciones que manipularan los datos de entrenamiento a fin de que una IA produzca resultados incoherentes? ¿Se puede quebrar la voluntad de una IA? ¿Y, llegado el caso, podría volverse una IA nihilista y abandonar todo el proyecto que se le haya asignado?

Ya hemos oído ejemplos de alucinación de ChatGPT3 cuando sus datos de entrenamiento no le proporcionan suficiente información sobre un tema. Esto se debe al intento surrealista por parte de la herramienta de rellenar sus lagunas de conocimiento con los datos disponibles, pero ¿y si la alucinación fuera un rasgo característico y no un error? ¿Se podría privar a una IA de suficiente información o entrenarla de tal manera que ignore sus datos de entrenamiento originales y aumenten sus episodios alucinatorios? Es algo parecido a poner a un preso en régimen de aislamiento. ¿Cuál sería el equivalente para la IA de tener las luces encendidas todo el día y perder la noción del tiempo y el espacio?

¿Por qué querría alguien entrenar de este modo a una IA? Parecería contradictorio, pero basta con ver la eficacia con la que los gobiernos –como en EEUU– han utilizado las redes sociales –y sus algoritmos de fidelización– para exacerbar la disidencia o una sociedad civil anémica, incapaz de ofrecer resistencia a la voluntad del Estado; algo que según el politólogo James Scott, es uno de los atributos definitorios de un estado fallido.

Es este último mecanismo de fracaso estatal el que más me preocupa. En el contexto actual, coinciden la necesidad creciente de tener certezas y el debilitamiento de las sociedades civiles. En su libro *La condición humana* (1958), Hannah Arendt argumentaba que la soledad era importante para ciudadanos democráticos, ya que daba a los individuos la capacidad de la contemplación, pero que el aislamiento era, en cambio, un camino hacia el totalitarismo. En el aislamiento, las personas se sienten tan abandonadas por sus conciudadanos que empiezan a cuestionarse a sí mismas y a todo lo que les rodea. Cuando los ciudadanos se sienten abandonados, son más vulnerables al totalitarismo. Arendt establecía así la diferencia con la tiranía, que es el gobierno del Estado motivado por el miedo. En la tiranía, la gente puede tener una vida privada que es incontrolable por parte del Estado, pero en el totalitarismo la ideología impregna a los ciudadanos de tal manera que no hay distinción entre la vida pública y la privada. ¿Puede utilizarse la IA para producir más aislamiento en los ciudadanos aumentando la incertidumbre del mundo que les rodea?

En tales circunstancias, cabría imaginar un pueblo inestable volviéndose hacia ideologías de la certeza que ven al Estado como un padre estricto dispuesto a imponer castigos para preservar la ley. En este contexto, una IA entrenada para hacer cumplir las normas adquiere mayor relevancia. ¿Tendremos quizás una IA dominadora y experta en las tablas de la ley para producir resultados morales correctos? Para llegar a ese punto sería preciso que nos aisláramos lo suficiente unos de otros y de nuestros propios sistemas de significado. Así pues, tenemos que recuperar el sentido de nosotros mismos, tal como somos, para evitar este destino y mantener una democracia sana y vivaz.