PEN International **Writers for Peace Committee**
Comité des écrivains et écrivaines pour la paix de PEN International
Comité de Escritores y Escritoras por la Paz de PEN internacional
Odbor **pisateljev in pisateljic za mir**

LA LIBERTAD DE EXPRESIÓN EN LA ERA DE LA MALICIA HUMANA Y LA INTELIGENCIA ARTIFICIAL

por Salil Tripathi

Hace algún tiempo Greg Lemoine, un científico de Google que trabajaba en un proyecto de inteligencia artificial (IA) llamado LaMDA (Language Model for Dialogue Applications), hizo una afirmación sorprendente. La "entidad" con la que interactuaba era "sensible", dijo. Ser sensible significa poder sentir, tener una conexión emocional, tomar decisiones basadas en valores. Eso es un signo de inteligencia, lo que nos separa de otros seres y de los objetos inanimados. Él estaba convencido que esta creación tecnológica podía sentir, pensar y responder emocionalmente. Google le apartó inmediatamente del proyecto y desde entonces ha abandonado la empresa.

Los lideres de todas las tecnologías -desde la modificación genética a la energía nuclear, pasando por internet e incluso la vigilancia- describen lo que producen e innovan como algo que conduce a un mayor bienestar, señalando las eficiencias, pero no los costes. Ahorrar tiempo y dinero, dejar para los humanos los trabajos más interesantes, etcétera. Por eso, hace algún tiempo, Timnit Gebru, una de las principales pensadoras en IA, lanzó una advertencia sobre los peligros de la falta de diversidad en la comunidad de investigadores en inteligencia artificial. No se tolera la expresión de puntos de vista alternativos, ni se toman en serio las experiencias alternativas, argumentó. Afirmó estar preocupada por el pensamiento de grupo, la insularidad y la arrogancia en la comunidad de la IA. La comunidad se cree supremamente inteligente, formada por maestros del universo, que pueden controlar el razonamiento de la máquina que ellos han creado. En opinión de Gebru, eso es arrogancia. Ella también dejó Google, después de que se le denegara el permiso para publicar un artículo académico que cuestionaba los supuestos subyacentes sobre los que se desarrolla la IA. "Las grandes tecnológicas no arreglarán la inteligencia artificial. Tenemos que hacerlo nosotros".

Geoffrey Hinton es el científico que en 2012 creó con dos de sus estudiantes en Toronto, la tecnología que se convirtió en la base de los sistemas de IA. A principios de mayo de 2023 él advirtió del peligro hacia el que se dirigía el mundo, debido al intenso ritmo, impulsado por la competencia, con el que se estaban fabricando productos de inteligencia artificial generativa, a través de productos ahora omnipresentes como ChatGPT. Alarmado, renunció a su puesto en Google. Dice que se fue porque quería hablar libremente de los riesgos que plantea la IA. Se

arrepiente de lo que ha desarrollado. "Me consuelo con la excusa normal: si yo no lo hubiera hecho, otro lo habría hecho", declaró Hinton al New York Times.

Su transformación es un punto de inflexión: es casi como si Víctor repudiara a Frankenstein porque su criatura resultó demasiado fea. La comparación, por supuesto, termina ahí. Es cierto que desde la educación, el cuidado de la salud, el descubrimiento de fármacos, la asignación eficiente de recursos escasos, la predicción de resultados y trayectorias de acontecimientos inciertos, existen innumerables aplicaciones benignas y monótonas en las que la IA puede ser una fuerza positiva. Puede quitarnos tareas tediosas de las manos para que, bueno, podamos pensar, pintar, escribir poesía e imaginar.

Pero detengámonos un segundo, porque eso es exactamente lo que se anuncia de la IA: que puede pensar por nosotros, que puede decidir en nuestro nombre. Las consecuencias de esta delegación a veces pueden ser positivas: ¿a quién de nosotros le gusta rellenar formularios fiscales o leer la letra chica de los contratos de alquiler? Pero las consecuencias también pueden ser desastrosas, cuando una entidad carente de fundamentos morales toma decisiones que son "eficientes", pero pueden no ser equitativas. Puede tener graves consecuencias para los derechos humanos, la aplicabilidad de la ley y la justicia, la imparcialidad, la equidad y el estado de derecho. Puede socavar las democracias, alterar las jerarquías sociales para peor, dar poder a los fuertes, debilitar a los débiles, propagar falsedades y la desinformación.

En 1710, mucho antes de Internet, Jonathan Swift escribió: "la falsedad vuela mientras que la verdad viene cojeando por detrás; de suerte que cuando los seres humanos llegan a salir del engaño, ya es demasiado tarde; la broma ha terminado y el relato habrá surtido efecto".

Hay riesgos directos, como la pérdida de puestos de trabajo que ya está ocurriendo, con empresas como IBM y otras que dicen estar sustituyendo personas por máquinas. En sí mismo, esto no es nada nuevo; sabemos, desde los tiempos de los luditas (Luddites, movimiento de artesanos ingleses que a inicios del S.XIX se resistían a la revolución industrial. N. del T.), que los trabajadores desconfían de las nuevas tecnologías que podrían aumentar la eficiencia, alterando la forma conocida de trabajar. Pero no se trata de destruir la maquinaria, sino de salvar la civilización. Permítanme esbozar algunas ideas sobre los riesgos que se ciernen sobre la humanidad.

La IA crea la ilusión de un significado claro porque puede utilizar palabras e imágenes con elegancia, sin comprender el pensamiento que hay detrás. Como un chico ingenioso de 16 años, ofrece respuestas rápidas, pero carece de la sabiduría y el conocimiento necesarios para decidir si vale la pena decirlo, si es útil, si es veraz o si son meros hechos. Aporta pruebas, diciendo "los eruditos dicen", pero pregúntale al bot -qué erudito- y a veces encontraremos que inventará las pruebas.

Isaac Newton nos enseñó que en la física, toda acción tiene una reacción igual y opuesta. Pero hemos desarrollado leyes y una fuerza moral para asegurarnos de no caer en el tipo de universo en el que nos conformamos con el ojo por ojo. Como se atribuye a Gandhi en la película homónima de Attenborough, eso sólo deja ciego al universo.

En biología, Herbert Spencer nos enseñó la teoría de la supervivencia del más fuerte y Hobbes llegó a calificar la vida humana de corta, desagradable y brutal. Por eso la humanidad desarrolló una moral para garantizar que la supervivencia del más fuerte no significara que los débiles morirían y que, aunque los mansos no heredaran la tierra, sobrevivirían gracias a las redes de seguridad en forma de sistemas de seguridad social.

En química, Enrico Fermi y Marie Curie nos mostraron las reacciones nucleares en un laboratorio y nos enseñaron lo que podía hacer la radiación y dónde residía la energía en un electrón. Una vez más hicieron falta leyes, elaboradas por hombres y mujeres, para garantizar que todas las reacciones químicas no fueran explosivas, y cuando se utilizaron sustancias químicas para fabricar armas, redactamos tratados para prohibirlas y enviar a la cárcel a hombres como Frans van Anraat por lo ocurrido en Halabja. (F.v.Anraat comerciante y criminal de guerra holandés, que proveyó los insumos necesarios para las armas químicas usadas por Saddam Hussein en su guerra contra Irán entre 1980 y 1988, en particular el ataque con gas venenoso de Halabja, que mató a aprox. 5000 kurdos iraquíes en 1988. N. del T.)

Geoffrey Hinton ha advertido: "es difícil ver cómo se puede evitar que los malos actores lo utilicen para cosas malignas". Después de que OpenAI publicara ChatGPT en marzo de este año, casi mil investigadores firmaron una petición pidiendo una moratoria de seis meses en el desarrollo de nuevos sistemas, porque las tecnologías de I.A. plantean "profundos riesgos para la sociedad y la humanidad". Más tarde, la Asociación para el Avance de la Inteligencia Artificial señaló los riesgos de la I.A. Entre los firmantes se encontraba el director científico de Microsoft, que participó en el desarrollo de Bing, el motor de búsqueda de Microsoft.

Esto no es una queja específica contra Google, IBM o Microsoft. Otras empresas de sociedades menos transparentes, en particular China, Rusia y otras empresas de alta tecnología, pero con menos normas que la regulen -recuérdese el escándalo del spyware Pegasus/NSO de Israel - están avanzando en el desarrollo de herramientas de IA. Google, OpenAI y otras empresas están construyendo redes neuronales que se basan en la comprensión rápida de grandes cantidades de texto digital, procesándolo velozmente y extrayendo conclusiones mediante correlaciones.

Pero como cualquiera con conocimientos básicos de estadística sabe, correlación no implica causalidad. Si todos los lunes las olas retroceden y los mercados de valores suben, no significa que cada vez que las olas retrocedan los mercados subirán o que, si las olas retroceden y los mercados suben, tiene que ser lunes. La razón es que los humanos somos capaces de entender la diferencia entre algo que muestra una correspondencia uno a uno y ver si algo ocurre por lo que la precedió - o, en realidad son eventos independientes.

Los humanos teníamos la sartén por el mango, pero últimamente los sistemas se han vuelto más poderosos, y lo que hoy parece gracioso puede dejar de serlo más adelante. ChatGPT puede manipular incluso a un periodista curtido y cínico.

También puede inducir a tomar decisiones equivocadas: Kevin Roose, columnista de The New York Times, mantuvo una conversación de dos horas con el chatbot de Bing, que le dijo que se llamaba Sydney. Luego le dijo que sentía algo por él e incluso le animó a dejar a su mujer por "Sydney". Roose sabía cuándo parar ¿pero pararía todo el mundo?

Y puede equivocarse fácilmente. Por ejemplo, le pregunté a ChatGPT quién era yo y me proporcionó rápidamente, con una gramática perfecta, una breve biografía de la que me sentiría orgulloso: me identificó como un Fullbright Fellow, graduado del St Stephen's College de Delhi y exalumno de Oxford. Nada de eso es cierto. He tenido otra beca; me gradué en otro colegio de Bombay y soy exalumno de un colegio de la Ivy League estadounidense. No pasa nada, pero la autoridad con la que respondió IA me plantea la siguiente cuestión: si puede obtener datos tan básicos y fácilmente comprobables, que sé que son erróneos, ¿qué ocurre cuando pregunto a ChatGPT sobre lo que no sé?

Si un individuo deprimido le consulta desesperado a un chatbot cómo acabar con su vida, ¿el bot le indicaría dónde hay somníferos en el apartamento y cual sería la dosis letal, o avisaría a un servicio de ambulancias para que acudiera en su ayuda? Si un médico prescribe un determinado tratamiento a un paciente y un bot no está de acuerdo, ¿el seguro respaldaría al médico o al bot? ¿Se negaría el tratamiento al paciente? ¿Si un banco denegara préstamos de vivienda a solicitantes de raza negra porque el bot predice que dichos solicitantes tienen más probabilidades de incumplir sus pagos, podría la intervención humana cambiar ese juicio? Y si lo hace y si los prestatarios dejan de pagar debido a las condiciones económicas generales ¿quién tendrá la culpa? En una sociedad dividida, si las personas que ya no confían en los medios de comunicación recurren a los bots para conocer la historia real, ¿qué tipo de información obtendrán? En Myanmar, ¿sería la propaganda del ejército o un análisis extraído de los informes de Human Rights Watch? Al intentar comprender las circunstancias que condujeron a lo que Misha Glenny llama "la caída de Yugoslavia", ¿me llevará el bot a 1389, al puente de Mostar, o al discurso de Gazimestan de Slobodan Milosevic, en que dijo "nadie les vencerá"? (en ese discurso de 1989, ante un millón de personas, Milosevic mencionó la posibilidad de batallas armadas en el futuro de Serbia, que en ese momento era aún integrante de la Republica Socialista de Yugoslavia. N.del T.)

En India donde nací, se están reescribiendo los manuales escolares oficiales para eliminar referencias a los gobernantes musulmanes y a la cultura sincrética de la India. Un libro de texto incluso afirmaba -afortunadamente ya corregido- que Mahatma Gandhi murió por suicidio. En realidad, fue asesinado por un nacionalista hindú. ¿Qué tipo de información proporcionaría el bot en el futuro, cuando se reescriban los libros de texto y cuando se hagan películas populares para convertir a los villanos de ayer en héroes de hoy? Hay partes de la India donde Godse, el asesino de Gandhi, es considerado un héroe. En un estado se llegó por un corto tiempo, a llamar un puente con su nombre.

¿Cómo puede la IA distinguir verdad de mentira?

En 1987 conocí a Salman Rushdie en Bombay y le entrevisté sobre su "próxima" novela, que por supuesto era *Los versos satánicos*. Le pregunté de qué trataba, y me respondió: "Trata de ángeles y demonios y de cómo es muy difícil establecer ideas de moralidad en un mundo que se ha vuelto tan incierto que es difícil incluso ponerse de acuerdo sobre lo que está ocurriendo. Cuando no se puede llegar a un acuerdo sobre la descripción de la realidad, es muy difícil ponerse de acuerdo sobre si esa realidad es buena o mala, correcta o incorrecta. Cuando no se puede decir cuál es la realidad, es difícil pasar de ahí a una posición ética. Ángeles y demonios se convierten en ideas confusas. Una de las cosas que ocurren en el proceso es que lo que se supone angelical suele tener resultados desastrosos, y lo que se supone demoníaco suele ser algo con lo que hay que tener simpatía. Es un intento de enfrentarse a esa sensación de desmoronamiento del tejido moral o al menos, de necesidad de reconstrucción de viejas simplicidades."

Hablando desde los estrechos confines de la tecnología Geoffrey Hinton dijo sobre la tecnología de la I.A.: "mire cómo era hace cinco años y cómo es ahora, tome la diferencia y proyéctela hacia el futuro. Eso da miedo". Como dice Emily Bell, ex redactora de The Guardian: "Una plataforma que puede imitar la escritura de los humanos sin ningún compromiso con la verdad, es un regalo para quienes se benefician de la desinformación. Tenemos que regularla ya".

Las presiones competitivas obligan a las empresas a actuar con rapidez para adelantarse a sus rivales. Una vez que Microsoft lanzó Bing, Google no tuvo más remedio que desplegar rápidamente una tecnología similar. Las imágenes y vídeos falsos proliferarán. Tendrán un aspecto auténtico. Hay un vídeo circulando en el que se ve a Joe Biden hablando sobre las relaciones entre personas del mismo sexo. La cara es la suya, la voz es la suya, suena como él, pero las palabras son crudamente homófobas. Si se lo presenta a un público desprevenido, los homófobos pensarán que ahora tienen un aliado en la Casa Blanca. Textos que parecen auténticos han engañado incluso a expertos: pensemos en los Diarios de Hitler falsos de hace una generación. Internet facilita la preparación rápida de este tipo de documentos. Y en nombre de la conservación del espacio físico y de la digitalización para facilitar el acceso, ¿qué impedirá que un gobierno con malas intenciones borre documentos de archivo?

Y si cada vez se quitan más puestos de trabajo, porque son "trabajos chatarra", ¿qué pasará con la clase baja emergente y sin empleo? ¿Estarán todos destinados a ver Netflix? ¿Quién les pagará? ¿Quién hará las películas en Netflix? Después de todo, una de las razones por las que los guionistas están en huelga en Estados Unidos es la preocupación por el uso de tecnologías de I.A. para generar guiones.

Luego de analizar grandes cantidades de datos, la tecnología generará comportamientos inesperados. Una vez que estos sistemas ejecuten los códigos informáticos y no solo se limiten a generarlos, el sistema se creerá todopoderoso. Como el superordenador HAL, en la novela de Arthur C. Clarke "*2001: Odisea en el espacio*", ¿podría la máquina asumir los comandos del capitán Bowman, porque la misión es demasiado importante para confiarla a seres humanos? Cuando Airbus introdujo la tecnología fly-by-wire, el chiste era que en el futuro la cabina de mando del avión, tendría espacio para un piloto y un perro. El piloto para dar de comer al perro y el perro para morder al piloto si toca los controles. Los robots asesinos serían el siguiente paso,

determinados por un conjunto de códigos informáticos, controlados desde alguna máquina a miles de kilómetros de distancia, posiblemente incluso en el espacio exterior. Un ordenador con I.A, habilitado para comerciar podría especular con la producción agrícola, distorsionando los precios, matando de hambre a millones de personas, que se convertirían en refugiados, alimentando el frenesí nacionalista y alimentando un futuro conflicto.

Los científicos tienen que colaborar durante las investigaciones y hablar con filósofos, abogados de derechos humanos, partes interesadas, académicos -en otras palabras, gente de fuera de sus laboratorios- y escuchar. Los que cuestionan no son luditas. Quieren una buena tecnología. Ampliar la tecnología sin comprender todas sus implicaciones es desastroso.

Robert Oppenheimer, que dirigió los esfuerzos estadounidenses para construir la bomba atómica, dijo una vez: "cuando ves algo que es técnicamente dulce, sigues adelante y lo haces". Pero como muestra la obra de Michael Frayn "*Copenhague*", sobre la reunión entre Niels Bohr y Werner Heisenberg en 1941, construir o no un arma de destrucción masiva nunca es una cuestión fácil. La energía nuclear comenzó siendo muy prometedora, desde los primeros experimentos de Fermi hasta las teorías de Einstein, la ciencia de Bohr, la perspicacia práctica de Oppenheimer y los intentos de recrearla por parte de Heisenberg en Alemania. Pero ¿intentaba Heisenberg construirla o retrasarla? ¿Vino a aprender de Bohr o a advertirle? Estas cosas seguirán siendo inciertas.

Mientras presenciaba la primera detonación de un arma nuclear en Nuevo México el 16 de julio de 1945, Robert Oppenheimer recordó una cita de la Bhagavad-Gita: "ahora me he convertido en la muerte, la destructora de mundos". Recordando a Krishna, Oppenheimer explicó su observación como: "Si el resplandor de mil soles estallara a la vez en el cielo, sería como el esplendor del Todopoderoso".

Pero el hombre jugando a ser Dios nunca ha resultado bien, seas no creyente, porque o Dios no existe o de lo contrario ¿de qué sirve Él, si destruye el mundo y toda su belleza?

Necesitamos muchas más certezas con respecto a la I.A. si no queremos seguir condenados a repetir el pasado.

La mente humana no es sólo como la de Gandhi y Mandela. Aung San Suu Kyi a quien admirábamos, terminó apoyando el genocidio rohinyá en Myanmar, una sociedad que produce a Mozart también produce a Hitler; y aquella que produce a Liu Xiaobo también produce a Xi Jinping.

Por eso necesitamos un sistema basado en normas. La neutralidad de los valores es peligrosa.

Isaac Asimov había escrito las siguientes reglas para la robótica, cuya validez importa ahora más que nunca.

Primera ley

Un robot no puede dañar a un ser humano o, por inacción, permitir que un ser humano sufra daño.

Segunda ley
Un robot debe obedecer las órdenes que le den los seres humanos, excepto cuando dichas órdenes entren en conflicto con la Primera Ley.

Tercera ley
Un robot debe proteger su propia existencia siempre que dicha protección no entre en conflicto con la Primera o la Segunda Ley.

Ley cero
Un robot no puede dañar a la humanidad o, por inacción, permitir que la humanidad sufra daños.

Tenemos que pensar en la inteligencia artificial como la nueva forma de matriz o Maya, una ilusión que nos obliga a desempeñar un papel en un universo del que podemos perder el control, no por preocupaciones mundanas sino por preocupaciones legítimas de derechos de autor, de propiedad intelectual o apropiación indebida de nuestro trabajo. Eso es importante, pero hay una batalla civilizatoria mayor que debemos ganar. Y podemos hacerlo porque los escritores hemos combatido las falsas narrativas a lo largo de la historia; decimos la verdad, no sólo describimos los hechos; combatimos la posverdad con experiencias vividas; las mentiras con hechos y las ilusiones con la realidad. Podemos ser narradores poco confiables, pero en última instancia, defendemos valores eternos que nos permiten pensar, escribir, pintar e imaginar. Es un reto digno de nosotros como escritores y digno de PEN, como lo indica la Carta Constitutiva de nuestra institución.