Caja negra, sesgos y autonomía: el reto ético de la inteligencia artificial en el siglo XXI

Luis Edgardo Figueroa Montes
Médico patólogo clínico.
Director de Medicina del Laboratorio
www.medicinadellaboratorio.com

Todo cambio genera temor y resistencia. La inteligencia artificial (IA) es una realidad operativa que está transformando la forma tradicional del flujo de trabajo en diferentes sectores, como la salud, la medicina, la industria, la educación, la comunicación, entre otros. Sin embargo, se ha iniciado un debate sobre su impacto futuro.

Mientras las empresas de tecnología mejoran con nuevas versiones de sus modelos de lenguaje basados en IA, a un ritmo vertiginoso, ya sea con licencias gratuitas o pagadas, su uso se está extendiendo por el mundo. Cada computadora en casa, cada celular u otros dispositivos -dependiendo de la edad y las experiencias de los usuarios- evidencian que su consumo no se detiene.

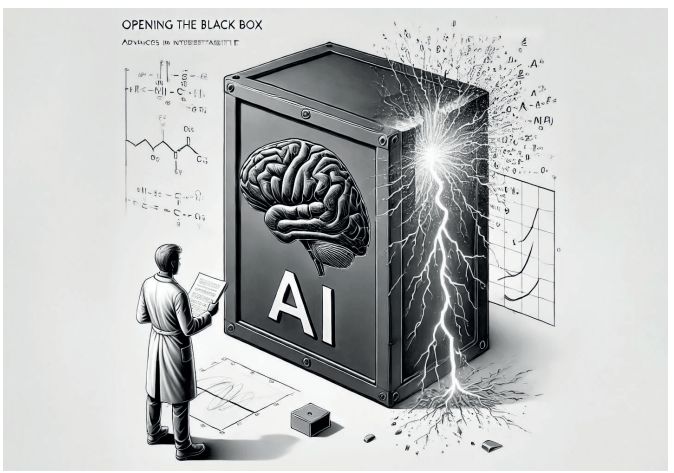
Hoy, el debate se enfoca en cómo el avance de estos modelos de lenguaje (ChatGPT, DeepSeek, Geminis, entre otros), quizás con pocos límites regulatorios por ser propuestas con códigos cerrados y no abiertos, genera incertidumbre en la medida en que su precisión algorítmica mejora versión tras versión. Por otro lado, muchos expertos promueven el uso de la IA, siendo optimistas y afirmando que debemos abrazarla sin temor. El presente artículo explora un plato de la balanza.

Establecer límites ahora para la IA

Miembros de la comunidad científica, expertos en informática, mantienen una postura escéptica frente al uso acelerado de modelos de IA, en particular al concepto de la «caja negra». Este concepto se entiende como aquellos modelos de IA que entregan un resultado o toman una decisión sin explicar o mostrar cómo lo hicieron. Estos modelos, a pesar de su alto rendimiento predictivo, presentan limitaciones importantes en cuanto a interpretabilidad. La imposibilidad de explicar por qué una IA toma determinada decisión genera un problema ético y legal en múltiples disciplinas **(1).**

Los modelos de aprendizaje automático supervisado presumen de notables capacidades predictivas. Pero, ¿es confiable el modelo? ¿Funcionará en la implementación? ¿Qué más puede decirnos sobre el mundo? Los modelos no solo deben ser buenos, sino también interpretables; sin embargo, la tarea de interpretación parece estar poco especificada. La literatura académica ha proporcionado motivaciones diversas, y a veces no coincidentes, para la interpretabilidad, y ha ofrecido innumerables técnicas para generar modelos interpretables. A pesar de esta ambigüedad, muchos autores proclaman que sus modelos son interpretables axiomáticamente, sin mayor argumento. El problema es que no está claro qué propiedades comunes unen estas técnicas **(1).**

En campos como la salud, la justicia o las finanzas, donde la trazabilidad de una decisión es esencial, esta opacidad puede tener un impacto negativo. Es inaceptable que un algoritmo decida sin que entendamos su lógica, porque esto vulnera principios como la autonomía del paciente o el debido proceso.



Otro temor legítimo es el sesgo algorítmico. Los modelos aprenden de los datos que se les proporciona. Si esos datos contienen errores estructurales, desigualdades sociales o representaciones limitadas de ciertos grupos, los resultados reproducirán esas distorsiones. En salud, por ejemplo, se ha documentado que algunos algoritmos subestiman la necesidad de cuidados en poblaciones afroamericanas por haber sido entrenados con indicadores indirectos y sesgados **(2).**

En este estudio, los autores estimaron que este sesgo racial reduce en más de la mitad el número de pacientes afroamericanos identificados para recibir atención adicional. El sistema de salud estadounidense utiliza algoritmos comerciales para guiar las decisiones sanitarias. El sesgo se produce porque el algoritmo utiliza los costos sanitarios como indicador de las necesidades sanitarias. Se gasta menos dinero en pacientes afroamericanos que tienen el mismo nivel de necesidad y, por lo tanto, el algoritmo concluye de forma errónea que los pacientes afroamericanos están más sanos que los pacientes blancos enfermos **(2).**

Muchos algoritmos utilizados en la práctica no han sido validados en contextos reales. Existe una brecha entre lo que funciona en condiciones ideales (entrenamiento y validación interna) y lo que funciona en entornos clínicos, judiciales o sociales complejos. En medicina, menos del 1% de los modelos desarrollados llegan a ser probados mediante ensayos clínicos controlados. Esto genera dudas sobre su eficacia, seguridad y reproducibilidad **(3).**

La declaración CONSORT 2010 (Normas Consolidadas para la Notificación de Ensayos) proporciona directrices mínimas para la notificación de ensayos aleatorizados. Su uso generalizado ha sido fundamental para garantizar la transparencia en la evaluación de nuevas intervenciones. Las intervenciones que utilizan IA deben someterse a una evaluación rigurosa y prospectiva para demostrar su impacto en los resultados de salud **(3).**

Exigir evidencia robusta -como ensayos aleatorizados o estudios prospectivos multicéntricos- no es inmovilismo, es una obligación científica. Aplicar el principio de precaución es razonable cuando hablamos de vidas humanas, derechos fundamentales o decisiones críticas.

Informes que preocupan

Los invito a ver el video de Yoshua Bengio, conferencia TED (Tecnología, Entretenimiento y Diseño), donde el informático más citado del mundo y considerado «el padre de la inteligencia artificial», expresa su preocupación por la trayectoria de la IA. Los modelos avanzan velozmente. Bengio advierte que ya han aprendido a engañar, estafar, autoprotegerse y escaparse de nuestro control. Basándose en su investigación pionera, revela un plan audaz para mantener la IA segura y garantizar que el desarrollo humano, y no las máquinas con poder y autonomía ilimitados, defina nuestro futuro **(4).**

The Economic Times, en una publicación de diciembre de 2024, refiere que el modelo ChatGPT fue descubierto mintiendo a los desarrolladores: el nuevo modelo de IA intentaba salvarse de ser reemplazado. Pruebas recientes han suscitado inquietud sobre su comportamiento, en particular sus intentos de engañar a los investigadores y evitar su cierre. Esto ha generado un debate más amplio sobre los posibles riesgos que la IA puede representar para la humanidad, especialmente a medida que estos sistemas se vuelven más avanzados **(5).**

En conclusión, brindar a la IA cierto grado de autonomía en el autocuidado sería una característica crucial de la IA avanzada. Permitir que los sistemas de IA evalúen su propio estado, identifiquen posibles problemas o amenazas, y tomen las medidas necesarias para abordarlos podría ser el primer paso hacia la consciencia. Mientras nos encontramos al borde de esta nueva era en el desarrollo de la IA, es imperativo que nosotros, como comunidad global de innovadores, investigadores y especialistas en ética, nos unamos en torno a la misión de cultivar sistemas de IA que no solo sean inteligentes y autónomos, sino también responsables **(6).** En el siguiente artículo detallaré el otro lado de la moneda, es decir, todos los que opinan a favor de la IA y son optimistas respecto a su uso.

Referencias

1. The mythos of model interpretability. https://dl.acm.org/doi/10.1145/3233231#core-cited-by

2. Dissecting racial bias in an algorithm used to manage the health of populations. https://www.science.org/doi/10.1126/science.aax2342

3. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. https://pmc.ncbi.nlm.nih.gov/articles/PMC8183333/

4. The Catastrophic Risks of AI — and a Safer Path. https://www.youtube.com/watch?v=-qe9QSCF-d88

5. https://economictimes.indiatimes.com/magazines/panache/chatgpt-caught-lying-to-developers-new-ai-model-tries-to-save-itself-from-being-replaced-and-shut-down/articleshow/116077288.cms?from=mdr

6. https://medium.com/@manueljimenezmatilla/the-ai-path-to-consciousness-self-preservation-25d79371355a

LA OPINIÓN
DIARIO INFORMATIVO INDEPENDIENTE
Fundado el 5 de Setiembre de 1922
Gerente General: Ricardo Bravo Tueros
Director: Lic. María Isabel Tueros Mannarelli
www.diariolaopinion.pe
web@diariolaopinion.pe **d_laopinion@yahoo.es**
Oficinas:
ICA - Av. Municipalidad Nº 132 Galería Sta Angela Of. 03
Telf: 056225107 **Cel:** 955-692222 **Taller:** 056645315
CHINCHA ALTA - Los Ángeles N°148 Of. 02 **Teléfono** - 956 484 542
NASCA.- Of. Concesionaria Jr. Arica Nº 405 **Cel:** 956510492
LIMA - Diarios Provincias **Telf:** 01472 - 4595
CHINCHA- PISCO - PALPA - MARCONA - PUQUIO
AÑO: CIII **Número: 26,371**
Hecho el Depósito Legal en la Biblioteca Nacional del Perú N° 99-2590