Cuando la inteligencia artificial entra al juzgado

MELISA LIZBETH SAAVEDRA GUTIÉRREZ

Imagínese que se encuentra en una audiencia donde su libertad depende de la decisión que se tome en la próxima hora, y que el juez a cargo de su caso no sólo consulta códigos legales o procedimientos estandarizados, sino que también se apoya en la inteligencia artificial que opera bajo términos opacos, o sea, de manera no totalmente cognoscible. Frente a él tiene una computadora que le muestra un número del uno al diez, el cual es generado por un algoritmo que evalúa si usted es peligroso o apto para volver a casa.

El uso de la IA se ha extendido al ámbito de la administración de justicia, sobre todo en el sistema penal de EE. UU., y su incorporación ha intensificado la falta de transparencia y el riesgo de que se reproduzcan sesgos en los datos, por ejemplo, contra sectores vulnerables. La dificultad para entender cómo la IA llega a una conclusión vuelve poco confiable el sistema legal y reduce la rendición de cuentas de las empresas desarrolladoras. Un caso reciente ocurrió en Florida, donde una inteligencia artificial falsificó un documento judicial sin consecuencias ni sanciones al proveedor del sistema.

Estas posibles consecuencias despiertan preguntas: ¿cómo podemos asegurar la transparencia en el uso de una IA?, ¿conforme a qué valores éticos se va a administrar la inteligencia artificial?, ¿de qué manera vamos a asegurarnos de que la IA los está aplicando óptimamente?

EL PROBLEMA DE LA CAJA NEGRA

Una de las inquietudes de esta forma de hacer justicia es la falta de un procedimiento y un enfoque que evalúe las dificultades de la opacidad epistémica. Los sistemas con IA no sólo ejecutan cálculos, sino que producen y legitiman conocimiento sobre el riesgo, la peligrosidad y la reincidencia, lo cual influye en decisiones que afectan la vida de las personas —en especial su libertad—. Por ello, debe analizarse qué tipo de conocimiento producen y bajo qué condiciones operan. En contraposición a esta IA de "caja negra" y secretista, una de "caja de cristal" permitiría verificar que los algoritmos no repitan prejuicios que perjudiquen a las minorías.

Si queremos evaluar con rigor los peligros éticos y los desafíos normativos que surgen de la incorporación de la inteligencia artificial en los sistemas de justicia, no basta, entonces, con señalar los riesgos abstractos: es imprescindible examinar su dimensión epistemológica. De este modo, una evaluación ética adecuada tiene que ponderar cómo se generan los datos, qué supuestos conceptuales estructuran los modelos, qué tipo de evidencia consideran relevante y bajo qué criterios se validan sus resultados.

ABRIR LA CAJA O CÓMO EVALUAR UNA IA

En 2024, Federica Russo, Eric Schliesser y Jean Wagemans —en su artículo "Connecting ethics and epistemology of AI"— pensaron estrategias para ayudar a crear una cultura de IA responsable. Los autores ofrecieron un marco teórico al que llamaron "epistemología-ética" que muestra cómo crear las condiciones para integrar valores éticos en todo el proceso de diseño, implementación, uso y evaluación de un sistema de IA, lo que posibilita que puedan ser verificados en una investigación y evaluación.

Russo y compañía señalan dos retos éticos importantes. Por un lado, en el proceso de programación algorítmica de la IA ocurren

sesgos relacionados a las medidas regulatorias de transparencia, lo que, a su vez, da paso al aumento de costos financieros en la mitigación. Y, por otro lado, al axiomatizar la imparcialidad en la IA —en términos de datos anónimos recogidos—, se pone en riesgo la privacidad, pues se da acceso a información potencialmente sensible.

Desde esta perspectiva, el modelo exige dos consideraciones: en primer lugar, en vez de confiar ciegamente en los resultados o la salida de la información dados por la IA, se debe poner a prueba la confianza de todo el proceso, desde el diseño inicial hasta el uso y la examinación de los resultados. En segundo lugar, hay que facilitar la evaluación de la fiabilidad o la integridad de los sistemas de inteligencia artificial, formulando una lista de preguntas críticas realizada por expertos.

Para lograr una internalización adecuada de valores éticos, los autores proponen la apertura del modelo —como si fuera una caja de cristal— con el fin de crear condiciones de transparencia, de modo que el código de la IA pueda reflejar principios y valores —como la participación, el acceso a la información y, en general, la democracia— en cada etapa. La implementación simultánea de mecanismos de evaluación ética y epistémica a lo largo de todo el ciclo de desarrollo y uso puede reducir riesgos sociales a largo plazo y consecuencias éticas en áreas tan delicadas como la justicia penal.

DATOS SUCIOS, DECISIONES INJUSTAS

En los últimos años, en las agencias de policía de Estados Unidos, se han desplegado sistemas de predicción para anticipar la actividad delictiva y asignar recursos legales en juicios para la evaluación de casos. En "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data", Richardson, Schultz y Crawford (2019) presentan estudios de caso de Chicago y Nueva Orleans en los que destacan el riesgo de perpetuar injusticias con predicciones tendenciosas. Estos sistemas se basan en datos producidos durante periodos de prácticas policiales defectuosas, racialmente sesgadas y en algunos casos ilegales, y que no han sido verificados previamente ni han sido recopilados con un método adecuado; en consecuencia se generan datos inexactos, "sucios".

Fotogramas de la película dirigida por Harun Farocki, *Prison images*, 2000.

En Chicago, el 56% de los hombres negros menores de treinta años tienen una puntuación considerada de riesgo en la Lista Estratégica de Sujetos, a pesar de que muchos nunca han sido arrestados. En el caso de Nueva Orleans, el Departamento de Justicia de los EE. UU. investigó dos veces a su policía. El resultado de dichas indagaciones mostró un patrón de uso de fuerza excesiva y de discriminación; los arrestos, por ejemplo, afectan en su mayoría a los residentes negros. En 2012, la ciudad contrató a la compañía privada de *software* Palantir con la intención de utilizar su servicio de análisis de datos para la prevención de la delincuencia; sin embargo, no hubo transparencia en los resultados por parte de la empresa y, además de evidenciarse prácticas abusivas, se reprodujeron prejuicios sociales. El sistema de IA

implementado identificó a los autores de delitos violentos como "abrumadoramente jóvenes, afroamericanos, varones con poca educación y subempleados".[1]

¿QUÉ VALORES DEBERÍA OBEDECER, ENTONCES, LA IA?

Para explicar cómo se podría establecer un código de conducta ética para la IA, tomaré como referencia una teoría normativa deontológica, la cual considera que a la hora de realizar u omitir acciones seguimos ciertos deberes morales que van más allá de las consecuencias de los actos.[2] La deontología kantiana, específicamente, estima que existen *deberes perfectos* que no admiten excepciones (por ejemplo, no mentir, no robar, no suicidarse); son deberes que Kant consideraba derivados de un imperativo categórico universal y necesario, es decir, válido independientemente de cualquier contingencia relativa a la experiencia o a casos particulares.

Este tipo de ética deontológica reconoce que los seres humanos son seres racionales con dignidad intrínseca, por lo cual no deben ser tratados como medios para un fin. La deontología contemporánea o moderada, basada en los principios de la asimetría de daño-beneficio, se plantea que las razones morales para no causar un daño son mucho más fuertes que las razones morales para producir un beneficio; es decir: es inaceptable matar a uno para salvar a otro, pero puede ser aceptable matar a uno para salvar a un millón de personas.[3]

Esta versión ética favorece actos de bajo riesgo moral, por ejemplo: un sistema de IA que detecte y elimine publicaciones potencialmente falsas sobre salud pública podría prevenir la desinformación masiva, pero también censurar información legítima de usuarios reales. El acto de bajo riesgo moral adecuado sería marcar el contenido como dudoso en lugar de eliminarlo. Otro caso ilustrativo es el de un sistema de IA que analice patrones de lenguaje en redes sociales para detectar posibles trastornos de salud mental con el objetivo de alertar a familiares o instituciones preventivas, evitando así una crisis del usuario. El acto de bajo riesgo moral correcto consistiría en no emitir ninguna alerta externa sin que la persona pida ayuda voluntariamente, limitando a la inteligencia artificial a ofrecer recursos de apoyo.[4]

Al integrar la teoría deontológica en la creación, implementación y evaluación de la IA se debe proponer, entonces, la existencia de deberes morales que hay que seguir independientemente de las consecuencias. Para esta ética, el respeto a la dignidad humana es un eje primordial, por lo que si la inteligencia artificial se programa con valores que especifiquen que las personas son sujetos de derechos, se reduce el sesgo.

Veamos un caso concreto. Si la IA detecta, mediante el reconocimiento facial, que alguien es sospechoso por su historial de navegación, detenerlo podría prevenir un delito. Sin embargo, ningún nivel de precisión estadística justifica privar a alguien de su libertad sin una prueba concreta, y el deber de respetar la presunción de inocencia es incondicional e independiente de delitos hipotéticos. La teoría deontológica rechazaría arrestar a la persona únicamente por las probabilidades futuras, pues se estaría tratando al sospechoso como medio para bajar la tasa de criminalidad y no como un fin en sí mismo, esto es, un hu-

1 Rashida Richardson *et al*., "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, And Justice", *New York University Law Review Online*, vol. 95, núm. 15, 05-2019, pp. 36-37.

2 William D'Alessandro, "Deontology and safe artificial intelligence", *Philosophical Studies*, vol. 182, 13-06-2024, pp. 1681-1704.

3 *Id.*

4 Masab A. Mansoor *et al*., "Early Detection of Mental Health Crises Through Artificial-Intelligence-Powered Social Media Analysis: A Prospective Observational Study, J*ournal of Personalized Medicine*, vol. 14, núm. 9, 2024; Soorya Balendra, "Meta's AI moderation and free speech: Ongoing challenges in the Global South", *Cambridge Forum on AI: Law and Governance*, vol. 1, 2025, pp. 1-19.

mano que tiene derechos, libertad, capacidad de decidir si va a delinquir o no y dignidad.

EL OBSTÁCULO: TRADUCIR LA ÉTICA AL CÓDIGO

Otro de los problemas de la inteligencia artificial es su aprendizaje mediante la automatización de datos, pues, como hemos comentado, sus redes neuronales artificiales, que reaccionan a un número determinado de entradas, replican comportamientos pasados y, en consecuencia, reproducen tendencias y un número considerable de errores.

Si bien hay que enseñarles de ética a los programadores, ahí no terminan los retos: ¿cómo traduciremos los valores deontológicos a axiomas que puedan entrar en la programación algorítmica? Sin una traducción adecuada, se compromete todo el enfoque. Aunque existen recomendaciones internacionales, como las de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), a menudo no hay un seguimiento real de su implementación.[5] Este problema limita nuestra propuesta deontológica y frena la incorporación de valores éticos complejos a la IA.

En particular, en el ámbito de la justicia penal, podemos resumir en dos los inconvenientes del uso de la inteligencia artificial: 1. La inflexibilidad por parte de la IA, pues los deberes perfectos vuelven compleja la forma en que se va a programar y aunque se lograse implementar una regla que le permita a la IA actuar para evitar un daño mayor, probablemente se radicalizarían sus acciones, y 2. La dificultad de enseñar y convertir en axiomas valores éticos complejos. Y es que el desafío central consiste en hacer programables, revisables y comprensibles los criterios éticos que deben orientarla. Sin esa traducción, incluso el modelo más riguroso en términos epistemológicos corre el riesgo de comprometer la dimensión normativa en su aplicación real. Para ello, entonces, hay que empezar por reducir la complejidad de los valores éticos.

5 OECD, "Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449), 3-05-2024; OECD, "Governing with Artificial Intelligence: State of Play and Way Forward in Core Government Functions", 18-09-2025.