MITOS Y REPRESENTACIONES DE LA INTELIGENCIA ARTIFICIAL

Editores

Gastón Becerra | Joaquín Mezzadra | Guillermo Movia

UFLO
UNIVERSIDAD

“La IA es objetiva”. Un análisis del sesgo en sistemas de IA

Ismael Ruiz

Los sesgos de la IA

La IA ya no es cosa de películas de ciencia ficción. Está en nuestros teléfonos, en las recomendaciones de series, en la preselección de candidatos para un trabajo e, incluso, en decisiones médicas de gran calado. Tan cotidiana, que resulta casi inevitable creer que es infalible, imparcial, pura como el cristal: una calculadora gigante que jamás se equivoca. Al fin y al cabo, los algoritmos son matemáticas y las matemáticas… ¿no están acaso libres de prejuicios?

Pero la realidad es que la IA no nace de la nada; la creamos los humanos. Y nosotros, con todas nuestras experiencias, historias y sí, también nuestros sesgos (conscientes o no), somos quienes la alimentamos con datos y diseñamos sus entrañas algorítmicas. Cuando los sistemas de IA se alimentan con datos defectuosos o limitados, el resultado es, previsiblemente, una salida sesgada tal y como exponen Ding et al. (2023). Es como si intentáramos enseñar a un niño solo con libros de historia que cuentan una única versión de los hechos: inevitablemente, el niño crecerá con una visión sesgada del mundo.

Esta omnipresencia ha llevado a una familiaridad tal con las experiencias impulsadas por la IA que resulta sencillo pasar por alto una realidad fundamental: todo sistema de IA alberga sesgos. La extendida adopción de dicha IA ha propiciado una concepción errónea sobre su objetividad, lo que puede llevar a subestimar las inclinaciones inherentes a estos sistemas. La aparente neutralidad que sugiere su base matemática ignora la influencia humana en la creación de datos y el diseño de algoritmos. La tarea de este mito es, precisamente, abordar esta idea equivocada.

La objetividad en el punto de mira

¿Por qué asociamos “máquina” con “objetividad”? Las raíces están en nuestra veneración por la ciencia y lo tecnológico: imaginamos a las máquinas como oráculos libres de pasiones. A esto se suma la falta de transparencia de muchos sistemas de “caja negra”, argumentos que ya expusieron Arrieta y el al. (2020) y

Kearns y Roth (2020) en sus respectivas investigaciones, cuya opacidad facilita presentar sus resultados como hechos incuestionables.

En el fondo, lo que se discute al hablar del mito de la IA objetiva es la tensión fundamental entre el ideal matemático de la computación y la realidad social de su creación y aplicación. El mito plantea un dualismo simplista: máquina = objetiva vs. humano = sesgado. Pero la IA es un sistema sociotécnico, una mezcla inseparable de código y contexto humano. Este mito bloquea una discusión crucial: la de la responsabilidad. Si la IA es "objetiva", ¿quién es responsable cuando toma una decisión injusta o discriminatoria? La culpa parece diluirse en el propio algoritmo. Al desmitificar la objetividad, abrimos preguntas necesarias: ¿cómo definimos y medimos la "justicia" en un algoritmo? ¿Quién debe rendir cuentas cuando un sistema de IA perjudica a un grupo? ¿Cómo podemos diseñar sistemas que no solo sean eficientes, sino también equitativos?

Conceptos como "objetividad", "inteligencia" o "progreso" son clave aquí y tienen múltiples sentidos. La "objetividad" en ciencia misma es un ideal complejo. En la IA, a menudo se confunde con la precisión estadística general, que puede ocultar un rendimiento muy desigual para diferentes grupos. Como bien explican Velander et al. (2024), la "inteligencia" de la IA es, hasta ahora, la capacidad de encontrar patrones en datos y hacer predicciones o generar contenido basado en ellos. Si los patrones en los datos reflejan prejuicios, la IA "inteligentemente" los replicará. El "progreso" en IA, medido solo por la capacidad o la eficiencia, no es verdadero progreso si deja a ciertos grupos atrás o amplifica desigualdades.

El sesgo, en este contexto, se define como el elemento clave que imposibilita que la IA sea objetiva, neutral o imparcial y, por ello, es de vital importancia abordar dicho concepto desde distintas perspectivas.

Fundamentos teóricos del sesgo

En el contexto de la inteligencia artificial, el término "sesgo" se define como un error sistemático en los procesos de toma de decisiones que conduce a resultados injustos o discriminatorios. Es crucial distinguir este concepto de los errores aleatorios o la varianza que pueden presentarse en los modelos de IA. Mientras que la varianza se refiere a la sensibilidad del modelo a las fluctuaciones en los datos de entrenamiento, el sesgo implica una desviación consistente de la verdad o la equidad. Comprender la naturaleza multifacética del sesgo requiere una taxonomía detallada de sus diferentes tipos, cada uno con sus propias características y orígenes.

Se han identificado diversos tipos de sesgo en la literatura científica reciente, cada uno de los cuales puede manifestarse en diferentes etapas del ciclo de vida de la IA:

Sesgo	Definición
De datos	Ocurre cuando los datos utilizados para entrenar los modelos son no representativos de la población real, están incompletos o reflejan prejuicios históricos presentes en la sociedad.
Algorítmico	Surge de las decisiones de diseño del algoritmo, las suposiciones inherentes que incorporan los desarrolladores o la manera en que el modelo aprende y procesa los datos.
De usuario	Es introducido por las personas que interactúan con los sistemas de IA, ya sea de forma consciente o inconsciente, al proporcionar datos sesgados o al utilizar el sistema de manera que refleje sus propios prejuicios.
De representación	Se presenta cuando ciertos grupos o características están subrepresentados o sobrerrepresentados en los datos de entrenamiento, lo que puede llevar a un rendimiento desigual del modelo.
Histórico	Se da cuando los modelos de IA aprenden y perpetúan los prejuicios que están presentes en los datos históricos utilizados para su entrenamiento.
De muestreo	Ocurre cuando los datos de entrenamiento no reflejan adecuadamente la diversidad de la población real a la que se aplicará el modelo.
De etiquetado	Se refiere a las inconsistencias o los prejuicios que pueden existir en el proceso de asignar etiquetas o categorías a los datos utilizados para el entrenamiento.
De agregación	Tiene lugar cuando la forma en que se combinan los datos oculta diferencias importantes que existen entre diferentes subgrupos dentro de la población.
De confirmación	Es la tendencia humana a favorecer la información que confirma las creencias preexistentes, lo que puede influir en la manera en que se interpretan los resultados generados por la IA.
De evaluación	Sucede cuando los modelos de IA se prueban utilizando conjuntos de datos que no son representativos del entorno en el que se desplegarán, lo que puede llevar a una sobreestimación de su precisión en situaciones reales.

De bucle de retroalimentación	Ocurre cuando un modelo de IA continúa aprendiendo de sus propias predicciones, lo que puede llevar a la amplificación de errores y sesgos con el tiempo.
Demográfico	Se refiere a los errores sistemáticos en los modelos que afectan de manera desproporcionada a grupos demográficos específicos, basándose en factores como la edad, el género o la etnia.
Institucional	Son los errores sistemáticos que surgen debido a las diferencias en las prácticas, los protocolos o los equipos utilizados en diferentes instituciones.
Temporal	Se refiere a los errores sistemáticos que aparecen con el tiempo, por ejemplo, debido a los cambios en la tecnología de imagenología médica, los protocolos o las características demográficas de los pacientes.

La categorización de estos sesgos proporciona una comprensión estructurada de la naturaleza polifacética del problema, destacando que el sesgo puede originarse en varios puntos del ciclo de vida de la IA. Además, es importante reconocer que estos diferentes tipos de sesgo no son mutuamente excluyentes y pueden interactuar entre sí, complicando aún más la tarea de identificarlos y mitigarlos.

Estos sesgos pueden infiltrarse en cada una de las etapas fundamentales del ciclo de vida de un sistema de IA. Durante la **recopilación de datos** se pueden introducir sesgos a través de la selección no representativa de la población objetivo o mediante el uso de datos históricos que ya contienen sesgos inherentes. Del mismo modo, en la fase de **preparación de datos**, el etiquetado subjetivo o inconsistente de la información, así como un manejo inadecuado de los datos faltantes, pueden ser fuentes significativas de sesgo.

También pueden darse en el **diseño del modelo** la elección de características que son irrelevantes para la tarea o que están correlacionadas con atributos sensibles (como la raza o el género), así como la optimización de métricas que no tienen en cuenta la equidad entre diferentes grupos pueden introducir o exacerbar sesgos.

Otra de las fases donde pueden surgir es durante el **entrenamiento del modelo**, donde el algoritmo aprende los patrones presentes en los datos, por lo que, si estos datos están sesgados, el modelo resultante también lo estará. En la etapa de **evaluación del modelo**, el uso de conjuntos de prueba que no son representativos de la población real o la aplicación de métricas de evaluación inadecuadas pueden ocultar la presencia de sesgos y llevar a una falsa sensación de confianza en el rendimiento del sistema.

Finalmente, en la **implementación y el uso** de los sistemas de IA, la interpretación sesgada de los resultados por parte de los usuarios o una confianza excesiva en las predicciones de la IA pueden perpetuar o incluso amplificar los sesgos existentes. Esta perspectiva del ciclo de vida subraya que el sesgo no es simplemente un problema de datos, sino una cuestión sistémica que requiere atención y medidas correctivas en cada etapa del desarrollo y la implementación de la IA.

Delegar y confiar

Una vez expuesto y comprendido el concepto de sesgo se debe poner en la palestra que el mito de la IA objetiva implica a una red compleja de actores sociales: los ingenieros y científicos de datos que construyen los modelos, las empresas que los desarrollan y despliegan, los usuarios que interactúan con ellos y las personas y comunidades cuyas vidas se ven afectadas por sus decisiones. Este mito parece hacer referencia principalmente a contextos donde la decisión automatizada busca ser "justa" o "eficiente", como la contratación, la concesión de créditos, la justicia penal o la atención médica. Pero si lo llevamos a otros contextos, como la educación, vemos que un algoritmo sesgado puede afectar la admisión universitaria, la asignación de becas o, incluso, la calificación de exámenes, perpetuando desigualdades desde temprana edad.

Del mito de la objetividad se desprende un mandato implícito: confiar en la máquina sin cuestionarla. Si la IA es objetiva, sus decisiones son "correctas" por definición. Esto supone una cosmovisión donde la tecnología es una fuerza neutral y benevolente, una especie de árbitro imparcial. Si todos creyéramos en este mito, podríamos terminar en una sociedad donde las desigualdades existentes se solidifican y amplifican por sistemas automatizados que nadie se atreve a auditar o corregir. Por el contrario, si todos pensáramos en contra del mito, seríamos usuarios y ciudadanos más críticos, exigiendo transparencia, rendición de cuentas y un diseño de IA que priorice la equidad. Como bien señalan los investigadores del campo educativo Rudolph, Tan y Tan (2023), la responsabilidad de usar la IA de manera crítica y ética está en nuestras manos.

Mirada sociológica

Desde la sociología, el mito de la IA objetiva nos habla de una sociedad marcada por la estratificación social y las desigualdades sistémicas. La IA, al entrenarse con datos históricos, absorbe y refleja los prejuicios y las estructuras de

poder presentes en esos datos. Conceptos como el "sesgo histórico" o el "sesgo de representación", explicados con anterioridad, son directamente sociológicos, mostrando cómo las dinámicas sociales se codifican en la tecnología.

Nuestra "imaginación sociológica" nos permite conectar la experiencia personal de, por ejemplo, no ser seleccionados para un trabajo o recibir un diagnóstico médico menos preciso, con el orden social más amplio. Quizás no fue solo mala suerte; quizás el algoritmo de contratación estaba sesgado contra nuestro género o edad, o el modelo de diagnóstico no fue entrenado con suficientes datos de personas con nuestras características demográficas. El mito naturaliza estos resultados sesgados como si fueran simplemente el producto de una evaluación "objetiva" de nuestros méritos o nuestro estado de salud, cuando en realidad son el resultado de procesos sociales y decisiones de diseño que introdujeron sesgos (Popenici y Kerr, 2017).

Así lo dicho, en la sociedad, que activamente trabaja para desmantelar las desigualdades históricas y sistémicas, la IA se debería desarrollar de manera diferente. Se debería priorizar la recopilación de datos diversos y representativos, así como el diseño de algoritmos que consideren la equidad, la realización de auditorías de sesgo regulares y la supervisión humana constante y significativa. De este modo, el mito de la IA objetiva perdería fuerza, ya que la tecnología misma reflejaría un compromiso consciente con la justicia social.

Información y datos

Popenici y Kerr (2017) defienden la idea de que la IA está intrínsecamente vinculada a los valores, creencias y sesgos de las culturas y los individuos que participan en su creación y utilización. De esta manera, las principales razones que identifican el por qué la IA no es neutral ni imparcial incluyen: sesgo en los datos de entrada y el diseño algorítmico; mecanismos opacos y falta de transparencia; influencia en la percepción y las elecciones; raíces históricas del concepto de "inteligencia"; sesgos cognitivos humanos, etc.

Como consecuencia, las manifestaciones de estos sesgos se han documentado en diversos ámbitos:

- En el ámbito legal, se contrasta la búsqueda de objetividad formal con el reconocimiento del realismo legal, que admite la influencia de factores psicológicos y sociales en las decisiones judiciales, un aspecto relevante al considerar la aplicación de IA en este campo, como bien investiga Goda (2022).

- En salud, la investigación desarrollada por Malerbi et al. (2023) destaca la necesidad de que los profesionales comprendan los sesgos de la IA y aboga por el acceso a conjuntos de datos diversos para mitigar la discriminación en modelos médicos.

- En el contexto académico, estudios recientes como, por ejemplo, el de Kamoun et al. (2024), señalan que los modelos de lenguaje como ChatGPT, aunque útiles, son propensos a errores y sesgos, y que la fiabilidad de sus respuestas a menudo se sobrestima. Esto plantea desafíos para la integridad académica y subraya la importancia de desmitificar la objetividad tecnológica.

A pesar de la solidez de la evidencia que refuta la objetividad de la IA, la investigación en este campo enfrenta ciertas limitaciones, tal y como exponen Jiang et al. (2022):

- La literatura sobre algunos aspectos, como la alfabetización en IA para educadores, aún es limitada.
- La comprensión pública de la IA a menudo se basa en información mediática, lo que puede generar conceptos erróneos.
- Las metodologías de investigación, como los cuestionarios, pueden no capturar la complejidad de las percepciones sobre la IA y los estudios a menudo se limitan a contextos geográficos o educativos específicos, lo que restringe la generalización de los hallazgos.
- La rápida evolución de la tecnología puede hacer que la investigación quede desactualizada con celeridad.
- Los propios sistemas de IA tienen limitaciones inherentes, como la capacidad de generar información incorrecta pero plausible.

Autores como Noble (2018), Broussard (2023) y O'Neil (2016) argumentan que los sesgos en la tecnología no son meros fallos, sino que a menudo son inherentes a su diseño y pueden exacerbar las desigualdades sociales.

Conclusiones

Entonces, ¿qué hay de verdad y mentira en el mito de la IA objetiva? La verdad es que la IA es una herramienta increíblemente poderosa y capaz de realizar tareas complejas a gran velocidad. La gran mentira es que esta capacidad la hace inherentemente neutral o imparcial. La IA es un espejo de los datos con los que se

entrena y de las decisiones humanas detrás de su diseño, y si esos datos y decisiones están sesgados, la IA también lo estará.

Desmontar este mito abre nuevas preguntas y discusiones fundamentales: ¿Cómo podemos construir sistemas de IA que no solo sean inteligentes, sino también justos y equitativos para todos? ¿Qué papel deben jugar la regulación y las políticas públicas para garantizar un desarrollo responsable de la IA? ¿Cómo educamos a la sociedad para interactuar críticamente con la IA, entendiendo sus limitaciones y potenciales sesgos?

Mi posición en esta discusión, basada en la evidencia científica, es clara: la IA no es objetiva por defecto. Es una tecnología con un potencial inmenso, pero también con el riesgo real de amplificar las injusticias sociales si no se aborda su desarrollo e implementación con una conciencia profunda de sus sesgos inherentes. Existe un claro consenso científico en rechazar la idea de la neutralidad de la IA y en destacar la urgencia de abordar el sesgo, la falta de transparencia y las implicaciones éticas para garantizar un desarrollo y uso responsable de esta tecnología. La clave está en reconocer que la IA es un producto sociotécnico y trabajar activamente, desde el diseño hasta la regulación y el uso, para construir sistemas que sirvan al bienestar de toda la sociedad.

Pero quizá la pregunta más urgente no sea si la IA puede ser justa, sino si estamos dispuestos a asumir la responsabilidad colectiva de construirla así. Porque en última instancia, el problema no está en la máquina, sino en las estructuras, prioridades y valores que decidimos programar en ella. Tal vez ha llegado el momento de dejar de preguntarnos qué puede hacer la IA por nosotros y empezar a preguntarnos qué tipo de sociedad queremos que refleje.

***

Para la realización de la escritura del presente mito utilicé diferentes IA: NotebookLM, Gemini y ChatGPT. En primer lugar, elaboré una búsqueda por Scopus de diferentes artículos científicos de calidad con los parámetros que me proporcionó ChatGPT y, en paralelo, le pedí a Gemini, mediante la Deep Research, la redacción de un informe sobre el mito tratado en este apartado. En segundo lugar, una vez filtrada la información y seleccionada, le pedí a ChatGPT un primer borrador del mito proporcionándole las fuentes relevantes, de calidad y con mayor impacto que busqué anteriormente. Además, le proporcioné una serie de condiciones de escritura: un tono formal pero desenfadado, inclusión de citas en apartados relevantes según la información proporcionada, una redacción llevadera, coherente e interesante, etc. En tercer lugar, a partir de dicho borrador fui escribiendo los distintos apartados del mito comprobando, a su vez, la veracidad de las fuentes citadas.

De este modo, la interacción con la IA ha resultado enriquecedora, puesto que fui yo mismo quien seleccionó las fuentes de calidad y, posteriormente, las comprobé para asegurarme de que eran fieles a aquello que se citaba. Además, me permitió la explicación de algunos pasajes de una manera más sencilla y comprensible.

Referencias

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R. y Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Broussard, M. (2023). *More than a glitch: Confronting race, gender, and ability bias in tech*. MIT Press.

Ding, L., Li, T., Jiang, S. y Gapud, A. (2023). Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education*, *20*(1), 1-18. https://doi.org/10.1186/s41239-023-00434-1

Goda, S.-L. (2022). Can we make all legal norms into legal syllogisms and why is that important in times of artificial intelligence? *Access to Justice in Eastern Europe*, *5*(1), 8-24. https://doi.org/10.33327/AJEE-18-5.1-a000095

Jiang, S., Nocera, A., Tatar, C., Yoder, M., Chao, J., Wiedemann, K., Finzer, W. y Rosé, C. (2022). An empirical analysis of high school students' practices of modelling with unstructured data. *British Journal of Educational Technology*, *53*(5), 1114-1133. https://doi.org/10.1111/bjet.13253

Kearns, M. y Roth, A. (2020). Ethical algorithm design. *SIGecom Exch.*, *18*(1), 31-36. https://doi.org/10.1145/3440959.3440966

Malerbi, K., Nakayama, F., Dychiao, G., Ribeiro, Z., Villanueva, C., Celi, A. y Regatieri, V. (2023). Digital Education for the Deployment of Artificial Intelligence in Health Care. *Journal of Medical Internet Research*, 25, 1-4. https://doi.org/10.2196/43333

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin.

Popenici, S. y Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*(1), 1-13. https://doi.org/10.1186/s41039-017-0062-8

Rudolph, J., Tan, S., y Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1), Article 1. https://doi.org/10.37074/jalt.2023.6.1.9