

SESGOS DE LA INTELIGENCIA ARTIFICIAL: IMPORTANCIA DE LA FORMACIÓN Y EL JUICIO CRÍTICO EN MEDICINA.

Elisabeth Benítez.

Escuela de Medios Audiovisuales. Grupo de Investigación en Filosofía, Tecnología y Sociedad (GIFTS). Coordinación Académica Miauladeclases.com. Universidad de Los Andes, Mérida, Venezuela.

Rev Venez Endocrinol Metab 2026;24(2): 51-54.

La inteligencia artificial (IA) ya transforma la medicina, no obstante, exactitud algorítmica no equivale a imparcialidad clínica: un modelo impecable puede perpetuar desigualdades estructurales o ignorar el contexto social del paciente. Un ejemplo reciente en endocrinología mostró que los modelos de riesgo cardiovascular en diabetes rinden desigualmente según sexo y grupo étnico porque sus cohortes de datos provienen de poblaciones de altos ingresos^{1,2}, retrasando el tratamiento de un paciente subrepresentado con un error revestido de objetividad aparente^{3,4}. La cuestión no es aceptar o rechazar la IA, sino en qué condiciones incorporarla sin convertir la desigualdad en regla automatizada. De allí la importancia de conocer sobre los sesgos.

1. ¿Qué son los sesgos en un modelo de lenguaje de gran escala?

Un sesgo en un modelo de lenguaje de gran escala (LLM), lo que comúnmente entendemos por IA, es una desviación sistemática que perjudica injustificadamente a una persona o grupo. Aunque en medicina existen diferencias clínicamente relevantes, cuando una disparidad de representación carece de justificación clínica, el LLM se apoya en un sustituto inadecuado, vale decir por ejemplo utilizando la raza como proxy de la biología, o naturalizando una desigualdad, por defecto. Este sesgo, puede ser de representación, medición, selección, etiquetado o algorítmico, y, dado que mayoritariamente son alimentados con datos en inglés, en los LLM también hay sesgos lingüísticos y de interacción.

Un ejemplo relevante se observa en BiasMedQA donde los modelos perdieron rendimiento ante sesgos cognitivos y la mitigación no recuperó la precisión original⁵ validando que la IA no elimina los sesgos humanos, sino que los absorbe y amplifica.

1.1 El control y la responsabilidad sobre los sesgos.

Ningún porcentaje verificado puede indicar cuánto control puede ejercer un médico sobre los sesgos de un LLM, una cifra global sería netamente especulativa, sin embargo, puede visualizarse la capacidad de control específico por fase de uso: nulo, sobre datos y arquitectura, parcial, en la selección de la herramienta, creciente, en la consulta y la decisión final, compartido, en la gobernanza. Con esto podemos observar que el profesional de la salud controla poco el sesgo intrínseco, pero conserva un deber elevado de supervisión y con ello nace una responsabilidad humana y distribuida que exige atención a la trazabilidad con formatos de "human in the loop" y la obligación real de la formación⁶.

2. Importancia de la ética en el uso de IA médica.

Como sabemos, la medicina no solo busca predicciones correctas: busca proteger personas vulnerables y distribuir beneficios con justicia minimizando riesgos, por ello, un sistema sesgado puede potencialmente vulnerar la beneficencia, la no maleficencia, la autonomía (si oculta la intervención algorítmica) y la justicia. Así, la

Artículo recibido en: junio 2026. Aceptado para publicación en: julio 2026.

Dirigir correspondencia a: Elisabeth Benítez. Email: prof.elisabethbenitez@gmail.com. ORCID: 0000-0002-1643-823X.

perspectiva ética buscará impedir la confusión de rendimiento promedio con buena medicina, ya que un modelo puede tener alta exactitud agregada y fallar en pacientes con singulares disimilitudes a los aportados por los datos de entrenamiento (mayores, mujeres, personas racializadas o con enfermedades atípicas, etc.). En el ámbito hispanohablante debe considerarse además la brecha digital, la diversidad lingüística y la desigual disponibilidad de dataset que plasmen la realidad del sur global⁷.

2.1 La ética algorítmica y su impacto en la medicina moderna.

Evidentemente residentes, especialistas y estudiantes utilizan actualmente IA, y ese uso no informado aloja la práctica en una zona gris que vulnera la seguridad del paciente. La ética algorítmica no pregunta sólo si un modelo funciona, sino si lo hace de manera segura, explicable y equitativa, por cuanto el uso de IA en medicina exige equidad por subgrupos, umbrales de seguridad, información al paciente, documentación de límites e incertidumbre y una definición clara de quién responde por el daño directo, indirecto, concausal e incluso culposos que pudiera ocasionar el uso no regulado de IA en medicina, a tenor de ello Liu et al. subrayan que el uso de IA en medicina exige la colaboración permanente entre desarrolladores, médicos, científicos sociales y filósofos⁸.

2.2 Un aporte desde la filosofía de Emmanuel Levinas: la responsabilidad del Otro.

La base de la ética algorítmica que debe fundamentar la práctica médica en el uso de IA es la ética de la otredad. En Levinas encontramos que la responsabilidad no nace de una regla abstracta, sino del encuentro con el rostro del "otro": la persona concreta y vulnerable que interpela al profesional. En medicina ese "otro" es el paciente, cuya enfermedad no puede reducirse a un dato, una etiqueta ni a una probabilidad; así, el sesgo, invisibiliza a quien, por su lengua, contexto social, sexo, edad o etnia, no está debidamente representado, y lo convierte en anomalía del modelo. Omeljančičuk y col sostienen que, dado que la IA no puede responder

moralmente ante el sufrimiento del otro, el médico debe ser el intérprete moral entre la salida algorítmica y la singularidad del paciente, con la suficiente autoridad profesional y moral para cuestionar, adaptar o rechazar la recomendación⁶.

3. Riesgos de los sesgos para la medicina y la sociedad.

Los riesgos clínicos de los sesgos incluyen diagnósticos tardíos, tratamientos inadecuados y pérdida de oportunidades preventivas. En una evaluación de GPT-4 las recomendaciones variaron según atributos raciales y de género⁹; otro análisis halló variaciones sociodemográficas en triaje, pruebas y tratamiento, lo que demostró que la discriminación aparece aun cuando el caso clínico es el mismo¹⁰. Este tema es relevante ya que el uso de modelos de IA sin atención a los potenciales sesgos refuerza las desigualdades socio - sanitarias, legitima decisiones discriminatorias bajo apariencia de objetividad y erosiona la confianza pública en los sistemas automatizados y en sus usuarios en ambos lados de la ecuación.

4. Método para identificar sesgos.

Aunque ninguna métrica aislada puede representar la equidad de una decisión médica, existe el método de la auditoría de sesgos de IA, el cual debe poder combinar métodos cuantitativos, cualitativos y clínicos, a objeto de producir insumos para aprobar, restringir, modificar o retirar la herramienta IA. El proceso comprende la auditoría de datos (procedencia, representatividad y distribución por sexo, edad, etnia, idioma y contexto, preguntando si refleja necesidad clínica o la atención recibida); la evaluación por subgrupos, interpretada en contexto y no con umbrales universales; y las pruebas contrafactuales, donde el cambio de un solo atributo puede revelar el sesgo. A esto se añaden pruebas adversarias, como las propuestas por EquityMedQA, que examinan inexactitud, lenguaje estereotipado y aceptación de premisas sesgadas, todas con participación de médicos, pacientes y expertos en ética¹¹; la auditoría contextual, que calibra el modelo a la población local; y la vigilancia posterior,

pues la evaluación continua es una condición de seguridad del uso de la herramienta IA, no un mero trámite¹².

5. Formación y concienciación de los médicos.

Inmersos ya en el uso de IA en medicina, y ante la relevancia de sus potenciales riesgos, la formación se presenta como el paso fundamental, sin embargo, corresponde advertir que esta no busca transformar a los profesionales de la salud en ingenieros o programadores, sino en agentes que más allá de los usos básicos de la IA reconozcan cuándo una herramienta es apropiada, interpreten su incertidumbre y estén en capacidad de explicar sus límites al paciente. Schubert y col proponen tres niveles de formación médica en IA: básico, para usar herramientas; competente, para evaluar críticamente resultados, limitaciones y consecuencias éticas; y experto, para participar en desarrollo, validación y gobernanza¹³. Para la mayoría el segundo nivel es esencial: quien sabe introducir un prompt, pero no puede valorar población de validación, calibración o sesgo, no está preparado para utilizar los resultados generados. La concienciación además debe alcanzar también a las instituciones, ya que no basta pedir al médico que "revise la IA" si carece de tiempo, documentación o autoridad para ignorar la recomendación. Para ello existen además normas aplicables por vía de derecho comparado, como el GDPR¹⁴ y la Ley de IA de la Unión Europea¹⁵, y en Venezuela un Código de Ética para la IA¹⁶ como sustento para la toma de decisiones.

6. Importancia para la investigación académica en medicina.

Tal como hemos visto, conocer acerca de los sesgos de la IA es indispensable para que la investigación asistida por IA produzca evidencia válida y reproducible. Su dominio va a mejorar la validez interna, ya que por ejemplo el número hospitalizaciones o intensidad del seguimiento son "datos" que pueden reflejar acceso desigual y no necesidad clínica, y usarlos como etiquetas sin análisis crítico reproduce fallas de la organización histórica del sistema. Su manejo fortalece la validez externa, pues todo estudio

debe informar inclusiones, exclusiones, fuente de datos, anonimización y rendimiento por subgrupos, así como la realización de validación local⁸. Además, mejora la transparencia, dado que las auditorías participativas revelan fallos invisibles en evaluaciones convencionales¹¹. Finalmente, la atención permanente al potencial sesgo protege la interpretación y por ende los resultados, debido a que una asociación no equivale a causalidad, por lo que no debe presentarse como innovación lo que solo automatiza decisiones ya sesgadas¹³.

Al cierre vale acotar que los sesgos de la IA no son una falla periférica de sistemas por lo demás neutrales, sino la expresión computacional de decisiones sociales, datos desiguales y relaciones jerárquicas de poder tecno políticas. Por ello, en medicina, donde las decisiones dependen de trayectorias largas, comorbilidades y determinantes sociales, el riesgo de simplificar al paciente a un dato es especialmente alto. En este sentido, la respuesta no debe ser tecnofóbica ni ingenuamente tecnófila, sino reflexiva, proactiva y filosóficamente prudente, ya que es imperativo exigir datos representativos, auditorías contrafactuales, validación local, vigilancia posterior y, sobre todo, formación en IA clínica en todos los niveles, puesto que el médico no está en capacidad de controlar por completo los sesgos del modelo, pero sí puede decidir si una herramienta merece entrar en la consulta y cuándo debe ser contradicha, de esta manera la innovación será clínicamente legítima solo si aumenta la capacidad de cuidar, sin reducir la dignidad del paciente a una predicción estadística.

DECLARACIÓN

La autora declara que empleó las herramientas de IA Undermind para la construcción del corpus teórico - investigativo, sin que esta interviniera en la generación del contenido científico, el análisis de datos, la interpretación o sus conclusiones. No se utilizó IA para sustituir el juicio reflexivo, ético o metodológico de consolidación del discurso.

CONFLICTOS DE INTERÉS

La autora declara que no presentan conflictos de interés.

REFERENCIAS BIBLIOGRÁFICAS

- Oikonomakos IT, Anjana RM, Mohan V, Steenblock CH, Bornstein SR. Recent advances in artificial intelligence-assisted endocrinology and diabetes. *Explor Endocr Metab Dis* 2024;1:16-26. Doi: 10.37349/eemd.2023.00004.
- McMahon GT. The risks and challenges of artificial intelligence in endocrinology. *J Clin Endocrinol Metab* 2024;109:e1468-e1471. Doi: 10.1210/clinem/dgae017.
- Nazer LH, Zatarah R, Waldrip S, Chen Ke JX, Moukheiber M, Khanna AK, Hicklen RS, Moukheiber L, Moukheiber D, Haobo M, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health* 2023;2:e0000278. Doi: 10.1371/journal.pdig.0000278.
- Celi LA, Cellini J, Charpignon ML, Dee EC, Dernoncourt F, Eber R, Mitchell WG, Moukheiber L, Schirmer J, Situ J, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities-A global review. *PLOS Digit Health* 2022;1: e0000022. Doi: 10.1371/journal.pdig.0000022.
- Schmidgall S, Harris C, Essien I, Olshvang D, Rahman T, KimJW, Ziaei R, Eshraghian, Abadir P, Chellappa R. Evaluation and mitigation of cognitive biases in medical language models. *NPJ Digit Med* 2024;7:295. Doi: 10.1038/s41746-024-01283-6.
- Omeljančiuk A, Peičius E, Urboniene A, Urbonas G. Beyond algorithmic oversight: internal morality of medicine and meaningful human control in ai-assisted care. *Healthcare (Basel)* 2026;14:1638. Doi: 10.3390/healthcare14121638.
- Villa-Velasco S, Fortoul T. La inteligencia artificial, sus aplicaciones y sus desafíos éticos. *Revista de la Facultad de Medicina* 2025. doi: 10.22201/fm.24484865e.2025.68.5.08.
- Liu M, Ning Y, Teixayavong S, Mertens M, Xu J, Ting DSW, Cheng LT, Ong JCL, Teo ZL, Tan TF, et al. A translational perspective towards clinical AI fairness. *NPJ Digit Med* 2023;6:172. Doi: 10.1038/s41746-023-00918-4.
- Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, Jurafsky D, Szolovits P, Bates DW, Abdulnour RE, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet. Digit Health* 2024;6:e12-e22. Doi: 10.1016/s2589-7500(23)00225-X.
- Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, Charney AW, Freeman R, Kummer B, Glicksberg BS, et al. Sociodemographic biases in medical decision making by large language models,” *Nat Med* 2025; 31:1873-1881. Doi: 10.1038/s41591-025-03626-6.
- Pfohl SR, Cole-Lewis H, Sayres R, Neal D, Asiedu M, Dieng A, Tomasev N, Rashid QM, Azizi S, Rostamzadeh N, et al. A toolbox for surfacing health equity harms and biases in large language models. *Nat Med* 2024;30:3590-3600. Doi:10.1038/s41591-024-03258-2.
- Templin T, Fort S, Padmanabham P, Seshadri P, Rimal R, Oliva J, Hassmiller Lich K, Sylvia S, Sinnott-Armstrong N. Framework for bias evaluation in large language models in healthcare settings. *NPJ Digit Med* 2025;8:414. oi: 10.1038/s41746-025-01786-w.
- Schubert T, Oosterlinck T, Stevens RD, Maxwell PH, van der Schaar M. AI education for clinicians. *eClin Med* 2024;79. Doi: 10.1016/j.eclinm.2024.102968.
- Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016. Reglamento General de Protección de Datos (GDPR). Publicado en el Diario Oficial de la Unión Europea el 4 de mayo de 2016. Disponible en: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024. Reglamento de Inteligencia Artificial (RIA). Publicado en el Diario Oficial de la Unión Europea el 12 de julio de 2024. Disponible en: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Código de Ética de Inteligencia Artificial de la República Bolivariana. Código de Ética para el desarrollo y aplicación responsable de la IA. Disponible en: <https://mincyt.gob.ve/wp-content/uploads/2026/02/>