Revista de Senología y Patología Mamaria 38 (2025) 100656

Sociedad Española de Senología y Patología Mamaria

Revista de Senología y patología mamaria
Journal of Senology and Breast Disease

www.elsevier.es/senologia

REVISIÓN

# Ética y filosofía de la inteligencia artificial aplicada a la enfermedad mamaria

**David Casacuberta**

*Departamento de Filosofía, Facultad de Letras, Universitat Autònoma de Barcelona, Barcelona, España*

Recibido el 8 de octubre de 2024; aceptado el 19 de noviembre de 2024
Disponible en Internet el 31 de diciembre de 2024

**PALABRAS CLAVE**
Aprendizaje automático en medicina;
Sesgos del aprendizaje automático;
Equidad de las aplicaciones de la inteligencia artificial (IA);
Ética de la inteligencia artificial (IA)

**Resumen** Este artículo aborda los desafíos éticos del uso de la inteligencia artificial en la detección de enfermedades mamarias, destacando que la fiabilidad de los algoritmos no es suficiente sin equidad. Basado en las ideas de Rawls sobre justicia, se discuten los sesgos en los datos y algoritmos que pueden afectar la precisión en poblaciones diversas. Además, se critica el dataísmo y se propone un enfoque que integre consideraciones éticas desde el inicio en el desarrollo de aplicaciones médicas de inteligencia artificial.


**KEYWORDS**
Machine learning in medicine;
Biases in machine learning;
Equity in artificial intelligence (AI) applications;
Ethics of artificial intelligence (AI)

**Ethics and philosophy of AI applied to breast pathology**

**Abstract** This article addresses the ethical challenges of using AI in breast pathology detection, highlighting that algorithmic reliability is not sufficient without fairness. Based on Rawls' ideas of fairness, it discusses biases in data and algorithms that may affect accuracy in different populations. Furthermore, it criticizes dataism and proposes an approach that integrates ethical considerations into the development of medical AI applications from the outset.


No es ningún secreto que la inteligencia artificial, especialmente aquella desarrollada desde los modelos de aprendizaje automático, está muy presente en nuestras vidas. Ningún aspecto de nuestra realidad cotidiana queda al margen de la capacidad de estos algoritmos. La inversión y

*Correo electrónico:* david.casacuberta@uab.cat

https://doi.org/10.1016/j.senol.2024.100656

desarrollo de algoritmos de IA en biomedicina no es ninguna excepción. De hecho, es uno de los campos donde más investigación y desarrollo podemos encontrar[1,2]. Un campo donde los algoritmos de aprendizaje automático son especialmente relevantes es la predicción: poder establecer un diagnóstico con cierta seguridad y con suficiente antelación como para poder responder de manera eficaz. De esos algoritmos esperamos exactitud y fiabilidad. Queremos que un algoritmo capaz de detectar un posible cáncer de mama, sea fiable y que sus predicciones se ajusten lo más posible a la realidad.

Pero, ¿es esto lo único que queremos? El uso de herramientas de predicción automática puede generar una serie de problemas cognitivos y éticos. Está, por ejemplo, la cuestión de garantizar la privacidad de los pacientes incluidos en las bases de datos para entrenar a los algoritmos. Está también la cuestión de la explicabilidad: ¿qué hace un profesional médico con una predicción que es simplemente un tanto por ciento, sin ninguna argumentación detrás? También está la confianza excesiva que genera una predicción automática, que puede hacer que el personal sanitario dude de su decisión original.

Básicamente, hay mucho más en juego en una predicción automática que la fiabilidad de la predicción. En particular, en este artículo quiero argumentar que hay algo más que deberíamos pedirle a un algoritmo que se vaya a usar en procesos de decisión que afecten al bienestar de las personas. Y esa otra cosa que esperamos de los algoritmos es que sus predicciones, además de fiables, sean *justas*.

Siguiendo las ideas de Rawls[3], entenderemos aquí «justicia» en el sentido de equidad, es decir, cuando aquellos valores y recursos que se consideran básicos para el bienestar de las personas están distribuidos de forma equitativa. Así pues, si la fiabilidad y la exactitud son valores que buscamos en los algoritmos de inteligencia artificial, establecer que esos algoritmos son justos significa asegurarnos de que esa exactitud está distribuida de forma equitativa y que no hay poblaciones discriminadas, para las que la fiabilidad del argumento no sea tan alta.

Consideremos el ejemplo del caso analizado[4] sobre un algoritmo que predecía la probabilidad de que una mancha en la piel fuera un melanoma o una mancha benigna. Originalmente los resultados eran muy prometedores: el algoritmo parecía incluso superar la capacidad de las personas expertas en dermatología. Sin embargo, poco tiempo después se observó que el algoritmo era mucho más fiable con pieles caucásicas que con pieles más oscuras de otras etnias.

Este sería un ejemplo perfecto de lo queremos exponer aquí: la fiabilidad no es suficiente. En términos globales el algoritmo era muy fiable, pero esa fiabilidad no estaba equitativamente distribuida entre poblaciones. Era un algoritmo fiable pero injusto.

Para comprender las implicaciones de esta distinción entre exacto e injusto en la práctica médica, debemos proceder de forma sistemática. El primer paso es reconocer que el problema existe.

Algunos sectores del mundo de la ingeniería creen que esta discusión es irrelevante, pues los datos en sí mismos no pueden estar sesgados: los datos en sí, por definición, son objetivos, nos dicen.

Una teoría o una hipótesis puede estar sesgada, pero no los datos sobre la que se sustenta.

De ahí se sigue, ya que un algoritmo no es más que matemáticas, que un algoritmo no puede estar sesgado, de la misma manera que no podemos decir que 12 X 2 = 24 está «sesgado» o es «discriminatorio». Esta es la posición argumentada, por ejemplo, por Pedro Domingos[5]. La posición extrema de esta propuesta se conoce como dataísmo, y en síntesis argumenta que para hacer ciencia en la actualidad los datos son más que suficientes. Las teorías son innecesarias. Es suficiente con introducir los datos en crudo en un ordenador, alimentar un algoritmo y buscar aquellas correlaciones que sean[6,7]. El dataísmo viene a ser la versión digital de una posición clásica en filosofía de la ciencia, el instrumentalismo, que defiende que una teoría científica no tiene valor de verdad más allá de aquello observable empíricamente y que las entidades teóricas no son más que instrumentos útiles para hacer predicciones, y que es irrelevante preguntarse por la verdad o falsedad de los constructos teóricos.

Si analizamos la cuestión con atención, veremos que esta posición no se sostiene. En primer lugar, es cuestionable que los datos sean realmente objetivos. Finalmente, una base de datos se construye a partir de una comprensión previa del problema por parte del equipo investigador sobre qué datos son relevantes y cuáles no. De manera que, ya de salida, los datos vienen «contaminados» por las suposiciones, las teorías o los prejuicios de las personas que desarrollan el proyecto, de manera que la idea de Chris Anderson de que estamos ante «el final de la teoría»[8] y que hipótesis o teorías ya no son necesarias no es más que una fantasía.

Por otra parte, si los datos que estamos recopilando ya están sesgados en origen, inevitablemente esos sesgos se trasladarán al algoritmo. Tal y como hemos comentado arriba con el ejemplo del algoritmo para detectar melanomas, si una etnia está infrarrepresentada en la base de datos porque en la vida real también lo está, esa infrarrepresentación va a afectar tanto a la exactitud como a la justicia de los resultados de ese algoritmo.

Algunos colectivos de desarrolladores de algoritmos de IA en medicina y otras áreas que implican el bienestar de las personas se agarran a ese segundo punto y le dan la vuelta, argumentando que el problema no son los algoritmos, sino la sociedad, que es discriminatoria. Los algoritmos no hacen más que reflejar las discriminaciones y segregaciones ya existentes en la sociedad. Es la sociedad la que ha de cambiar, no los algoritmos.

Aunque formalmente sea verdad, es inaceptable como justificación. ¿Se imaginan a un juez decidiendo que una persona ha de ir a la cárcel por tráfico de drogas y que su único argumento es que esa persona es afrodescendiente, y que la sociedad ya discrimina a esa etnia considerando que son todos criminales, y él se limita a seguir esa tendencia de la sociedad? Claro que no. Por muy humano que sea el juez, finalmente lo que esperamos de él es justicia, no que repita los prejuicios existentes en la sociedad.

Otra excusa que ciertos desarrolladores de algoritmos usan para evitar discutir las repercusiones éticas de su trabajo es que «las personas también se equivocan». Es decir, no podemos esperar que los algoritmos sean exactos al 100% o que no cometan ningún tipo de sesgo ético, ya que las personas también cometemos errores. Eso es cierto, sin duda, pero de nuevo como excusa es muy pobre. Si una doctora se equivoca no se encogerá de hombros y murmurará

«todas nos equivocamos», sino que procederá a intentar enmendar su error y, si ya no se puede, se aplicará para no cometerlo otra vez. La misma actitud hemos de esperar de una persona que desarrolla algoritmos de IA.

De todas formas, hay una diferencia muy relevante entre errores humanos y errores de máquinas. Los errores humanos se distribuyen en multitud de categorías y la mayoría de las veces los errores de unos quedan compensados por los errores de otros[9]. Un influencer en las redes sociales profundamente misógino puede quedar compensado por una alcaldesa activista fuertemente feminista, un juez de origen asiático cancela a otro con tendencias racistas, etc. Cada persona tiene sus propios prejuicios y sesgos, diferentes de los demás.

En cambio, los errores algorítmicos escalan. Si a diferentes instancias de un mismo tipo de algoritmo de aprendizaje automático le damos la misma base de datos, o simplemente una base de datos en los que el grueso de los datos coincide, los tipos de errores que cometerá serán siempre los mismos. Si luego, esos resultados se aplican en el mundo real y se recopilan los nuevos datos para alimentar nuevos algoritmos, los mismos errores se seguirán transmitiendo y acabarán teniendo cada vez un peso mayor en el algoritmo.

Imaginemos una situación en la que diagnosticar el cáncer de mama se hace de manera rutinaria utilizando algoritmos. Imaginemos también que esa base, al estilo de la de los melanomas que comentábamos antes, tiene un claro sesgo poblacional, y que la mayoría de las mujeres pertenecientes a esa base de datos son de origen caucásico. Si las predicciones de ese algoritmo se usan directamente, sin un cuestionamiento previo y son todo el criterio para futuras intervenciones y acciones, y luego los resultados de esas acciones futuras se usan para volver a entrenar el algoritmo, el sesgo a favor de las mujeres caucásicas será cada vez más presente, se irá multiplicando cada vez más, pues los errores no se cancelan, sino que se expanden[10].

Una vez hemos aceptado que el problema es real, quienes desarrollan algoritmos para ser usados en un contexto médico han de asegurarse no solo de la exactitud de sus algoritmos, sino también de su equidad.

¿Qué elementos hay que tener en cuenta? El primero y más central es comprometerse de manera sincera con la equidad en los algoritmos. Dejar de pensar que hacer que la valoración y el seguimiento ético no es más que una carga con la que cumplir una vez terminada la aplicación e ir marcando como hechos los requerimientos que nos pone la agencia de protección de datos. En su lugar, hemos de ser proactivos y diseñar desde el principio nuestra aplicación pensando en las implicaciones éticas y sociales de nuestro algoritmo. Es decir, las consideraciones éticas han de acompañar el proyecto desde el principio, introduciendo requerimientos éticos en todas las fases de desarrollo del proyecto[11]. Cuando definimos las diferentes variables que consideramos relevantes para nuestra investigación, pensemos en los posibles sesgos que pueden aparecer. Al recopilar la base de datos con la que entrenaremos nuestro algoritmo, revisemos posibles sesgos ocultos en esa base de datos. Si usamos una base de datos ajena, inquiramos cómo se han obtenido esos datos y hasta qué punto son representativos para la tarea que vamos a modelar con el algoritmo.

En ese proceso es importante entender la naturaleza de los sesgos y de qué manera pueden afectar a la equidad y exactitud de nuestro algoritmo. Una primera cosa a tener en cuenta es que «sesgo» es una palabra polisémica, que significa cosas diferentes en función del contexto. Tenemos así, el sesgo estadístico, que es simplemente la diferencia entre valor resultante y el valor esperado cuando hacemos predicciones. Es el concepto de sesgo que una persona experta en estadística entiende y usa de manera coherente e informada.

Luego tenemos el «sesgo en los datos», que resulta cuando nuestra base de datos no refleja a la población que supuestamente está capturando. Así, la ya mencionada base de datos sobre melanomas estaba sesgada, pues no representaba a la población que quería capturar.

Este sesgo puede estar presente directamente en la base de datos de salida, o puede aparecer cuando el equipo desarrollador toma la base de datos y la rehace, recogiendo solo las entradas que tienen los datos que se consideran relevantes, sin darse cuenta de que están eliminando una parte relevante de la población a estudiar. Es lo que conocemos como sesgo por selección. Imagine, por ejemplo, una base de datos que contiene información de diferentes hospitales. En uno de esos hospitales se hicieron una cantidad relevante de análisis clínicos que incluían una serie de datos sobre los pacientes que en los otros hospitales no se incluían. Imaginemos que ese hospital es privado y de precios elevados. Si el equipo que recorta la base de datos usa esos datos de análisis clínicos para seleccionar qué pacientes entran en la base de datos y cuáles no, acabará con una base de datos donde solo hay personas con un gran poder adquisitivo, con lo que no representará fielmente a la población a investigar.

El sesgo por selección puede aparecer también debido a un sesgo previo de la confirmación. Sucede cuando los investigadores están ya convencidos de cuál es la relación entre las diferentes variables y el resultado final y, consciente o inconscientemente, practican el *cherry picking* y solo incluyen aquellos datos que confirman su hipótesis de salida, descartando evidencia refutadora como «artefactos», «errores de medición», etc.

Un sesgo poblacional puede llegar también en función del tipo de instrumentos de medida que usemos, pues no representan de la misma forma a diferentes poblaciones. Un ejemplo clásico son los tests de inteligencia, que se basan sobre todo en la comprensión lectora. Personas de extracción social baja, menos expuestos a la lectura, tienden a puntuar más bajo que personas de clase de media, pero no porque sean menos inteligentes, sino porque no acaban de entender algunas expresiones lingüísticas, o los ejemplos que se les pone no forman parte de su experiencia diaria.

El país productor del algoritmo puede generar también un tipo de sesgo, el sesgo geográfico. Este sesgo tiene lugar cuando se entrena un algoritmo con datos de un país específico y luego se usa en otro, con otra población diferente. Será un sesgo cada vez más común a medida que aplicaciones de IA desarrolladas en China se ofrecen a países europeos sin ser conscientes de que las características médicas de la población china diferirán de la española en muchos aspectos relevantes.

Otro sesgo preocupante, que en este caso se da en las mentes de los profesionales sanitarios, es el sesgo de la automatización. Este sesgo tiene lugar cuando, frente a una predicción equivocada, el personal sanitario decide confiar más en la predicción, por su supuesta objetividad, que en su propio criterio. Tal y como se argumenta en Dratsch et al.[12], con un resultado incorrecto de un programa para analizar mamografías, los radiólogos tendían a confiar más en el programa que en su propio criterio.

El resultado final es que la habilidad para diagnosticar correctamente, radiólogos con poca experiencia o experiencia moderada, caía del 80% a un mero 25%. Los radiólogos con mucha experiencia tampoco eran inmunes, reduciéndose su capacidad de hacer un diagnóstico correcto del 80 al 45%.

Finalmente, el ejemplo que poníamos de lo que podría suceder si para predecir cánceres de mama se reutilizarán los datos generados por los algoritmos para entrenar a nuevas versiones del algoritmo, tendríamos lo que se conoce como sesgo de retroalimentación *(feedback loop)* y es uno de los más preocupantes, pues tiende a exacerbar sesgos ya presentes en la sociedad y convertirlos en inevitables.

La inteligencia artificial ha venido para quedarse. Sería ridículo desaprovechar todas las oportunidades que nos ofrece. Sin embargo, es necesario ser cautos. Hemos de evitar todo el *hype* que hay últimamente sobre la IA y cómo va resolver todos nuestros problemas. En este proceso de analizar qué aplicaciones tienen sentido en medicina y cuáles no, es muy importante establecer la fiabilidad de las aplicaciones de la IA en patología mamaria. Pero no tenemos que quedarnos ahí.

La exactitud, aunque necesaria, no es suficiente. Necesitamos ir más allá y preguntarnos por la equidad de esas aplicaciones. ¿Respetan los derechos humanos básicos? ¿Están sesgadas hacia unos colectivos olvidando otros? Y no es una pregunta que hagamos al final para quedar bien, sino ha de ser un principio epistémico que guíe todo el desarrollo de aplicaciones de inteligencia artificial en medicina.

## Consentimiento informado

David Casacuberta Sevilla, en calidad de autor del artículo, cuenta con el consentimiento de los pacientes para su publicación.

## Financiación

Este trabajo ha sido financiado por el Ministerio de Ciencia, Innovación y Universidades dentro del Subprograma Estatal de Generación del Conocimiento a través del proyecto de investigación PID2023-148517NB-100. Este trabajo parte de la red de investigación consolidada (GEHUCT), reconocida y financiada por la Generalitat de Catalunya, referencia 2021 SGR 00517.

## Conflicto de intereses

No hay conflictos de intereses.

## Bibliografía

1. Savage N. Tapping into the Drug Discovery Potential of AI Nature com; 2021. https://www.nature.com/articles/d43747-021-00045-7. AI;Nature:com.
2. Devereson A, Macak M, Nagra N, Idoux E. AI in Biopharma Research: A Time to Focus and McKinsey; Company. Disponible en: https://wwwmckinseycom/industries/life-sciences/our-insights/ai-inbiopharma-research-a-time-to-focus-and-scale, 2022 última visita 9-09-2024.
3. John Rawls. A Theory of Justice Revised Edition Harvard University Press Justice; Revised:Edition. Harvard University Press; 1999.
4. Groh M, Badri O, Daneshjou R, Koochek A, Harris C, Soenksen LR, Picard R. Deep learningaided skin;tones. Nat Med. 2024;302:573–83.
5. Domingos P. We must Stop Militant Liberals from Politicizing Artificial Intelligence The Spectator December 22 2020 Disponible en: https://spectatorus/militant-liberals-politicizing-artificial-intelligence, 2020 última visita 9-09-2024.
6. Van Dijck J. Datafication dataism and dataveillance: Big Data between scientific paradigm and ideology. Surveill Soc. 2014;122:197–208.
7. Larsen M. Toward a dataist future: Tracing Scandinavian posthumanism in Real Humans AI SOCIETY 1–13. Real:Humans AI SOCIETY; 2023. p. 1–13.
8. Anderson C. The end of theory: the data deluge makes the scientific method obsolete. Wired Mag. 2008;167 method; obsolete:Wired magazine 167 16-07.
9. James Surowiecki. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Societies Nations Anchor Books; 2005.
10. Guersenzvaig A, Casacuberta D. La quimera de la objetividad algorítmica: dificultades del aprendizaje automático en el desarrollo salud. IUS:Sci. 2022;81:35–56.
11. Casacuberta D, Guersenzvaig A, Moyano-Fernández C. Justificatory explanations in machine learning: for increased transparency through documenting how engineering;decisions. AI & Soc. 2024;391:279–93.
12. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, Pinto dos Santos D. On reader:performance. Radiology. 2023;3074, e222176.