DOI: 10.32440/ar.2026.143.01.rev05
Enviado: 14.02.26 | Revisado: 22.02.26 | Aceptado: 28.03.26

REVISIÓN

INTELIGENCIA ARTIFICIAL Y MEDICINA: DIEZ LECCIONES APRENDIDAS (Y OLVIDADAS): 1970-2026

ARTIFICIAL INTELLIGENCE AND MEDICINE-TEN LESSONS LEARNED (AND FORGOTTEN): 1970-2026

Víctor Maojo[1,2]; Casimir A. Kulikowski[3,4]

1. *Catedrático de Inteligencia Artificial, Universidad Politécnica de Madrid.*
2. *Académico Correspondiente de la Real Academia Nacional de Medicina de España.*
3. *Board of Governors Professor of Computer Science, Rutgers University, EE.UU.*
4. *Miembro de la National Academy of Medicine, EE.UU.*

Palabras clave:
Inteligencia Artificial; Medicina; Sistemas expertos; Aprendizaje automático; Aprendizaje profundo. IA generativa.

Keywords:
Artificial Intelligence; Medicine; Expert systems; Machine Learning; Deep Learning; Generative AI.

Resumen

El auge de la Inteligencia Artificial (IA) en los últimos años, debido a la aparición de sistemas de la llamada IA generativa, ha causado un enorme impacto científico, tecnológico y social, con sustanciales resultados y promesas en todas las áreas de la medicina. En este artículo se analiza la situación actual de la IA en medicina, comparando diversos temas y áreas con experiencias pasadas vividas por los investigadores en IA desde 1970. Sugerimos diez lecciones aprendidas de los éxitos y fracasos de estos años, y cómo algunas de estas deficiencias se repiten ahora de manera similar, lo que podría retrasar las promesas de cambio en la medicina del futuro.

Abstract

The rise of Artificial Intelligence (AI) in recent years, driven by the emergence of generative AI systems, has had an enormous scientific, technological, and social impact, yielding substantial results and promising outcomes across all areas of medicine. This article analyzes the current state of AI in medicine, comparing various topics and areas with past experiences of AI researchers since 1970. We suggest ten lessons learned from the successes and failures of these years, and how some of these shortcomings are being repeated now in similar ways, potentially delaying of the promises of change in future medicine.

INTRODUCCIÓN

La Inteligencia Artificial (IA) es una disciplina científica y tecnológica, tradicionalmente dividida en dos subcampos conceptuales: (1) simbólica y (2) conexionista. Desde una perspectiva práctica, podemos considerar dos etapas fundamentales de la IA en medicina. La primera etapa, dominada por los sistemas basados en el conocimiento, abarcó desde 1970 hasta aproximadamente mediados de la década de 1990. La segunda etapa, dominada por los sistemas basados en datos, comenzó a mediados de la década de 1990 y continúa hasta la actualidad con sistemas desarrollados bajo el paraguas de términos como "machine learning", "deep learning"(a partir de ahora, aprendizaje automático y aprendizaje profundo, respectivamente) e IA generativa.

Las aplicaciones de la IA en medicina que surgieron a partir de 1970 fueron consecuencia directa del análisis exhaustivo previo de los procesos cognitivos que subyacen al razonamiento y la toma de decisiones médicas. Investigadores como Ledley y Lusted (1), Pauker y Gorry (2), Gorry (3), Warner (4), Tversky y Kahneman (5), Feinstein (6), Simon (7) y otros investigaron temas como: las heurísticas utilizadas por los médicos en su razonamiento; la toma de decisiones en condiciones de riesgo e incertidumbre; el uso de varios tipos de razonamiento lógico para diagnóstico y terapia médica; el manejo de pruebas diagnósticas; el uso del teorema de Bayes para la estimación de probabilidad; o el análisis estadístico de bases de datos de pacientes, entre otros.

Los primeros sistemas de IA en medicina tuvieron éxito clínico limitado, e incluso se los ha llamado, despectivamente, "depósito de chatarra" (8); pero esos primeros resultados llevaron a éxitos como el desarrollo de sistemas de apoyo a la toma de decisiones médicas, métodos y técnicas subyacentes a las ontologías, terminologías biomédicas o las historias clínicas electrónicas, o técnicas de recuperación de

Autor para la correspondencia
Víctor Maojo
Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingenieros Informáticos. Campus de Montegancedo UPM
Tlf.:+34 910 672 898 | E-Mail: vmaojo@gmail.com

información en sistemas como Pubmed, entre otros. Estos sistemas pioneros tuvieron deficiencias y errores, en una época en la que no se disponía de experiencia previa y la investigación comenzaba desde cero. Sorprendentemente, muchas de estas mismas deficiencias y conceptos erróneos parecen repetirse hoy, en aplicaciones de aprendizaje profundo e IA generativa. En este artículo, inspirados especialmente en nuestra propia experiencia en investigación avanzada en IA en medicina desde las décadas de 1960 (CK) y 1980 (VM), queremos señalar una serie de lecciones aprendidas (y a veces olvidadas) de la IA en medicina.

LA IA NO ES NUEVA - Y LO QUE ESTO IMPLICA -.

La aparición de ChatGPT en 2022 ha provocado un aumento enorme en el número de usuarios de IA. En este contexto, muchas personas desconocen la historia de la IA, que surgió oficialmente como disciplina en 1956 en una reunión celebrada en el Dartmouth College por diez pioneros ahora legendarios (9).

Lo ocurrido en esta reunión no puede entenderse sin la investigación sobre matemáticas y lógica del siglo XX, el desarrollo previo de la Cibernética por Wiener o la conferencia Macy de Ciencias Cognitivas, celebrada ese mismo verano de 1956; sin embargo, pocos profesionales (incluso universitarios) parecen preocuparse por la historia de la IA y su relevancia para los desafíos actuales, ignorando que, sin dicha información, no es posible comprender plenamente las complejidades de la IA.

El aprendizaje en IA tiene un componente estadístico clave y, por esta razón no podemos afirmar que esté libre de errores (10). Como ha descrito Jelinek, pionero del procesamiento del lenguaje natural ("natural language processing" o NLP, nombre y acrónimo en inglés, usado comúnmente), la estadística ha dominado ese área por razones prácticas (11). En esencia, los sistemas de IA generativa son sistemas estadísticamente eficientes que a menudo producen resultados espectaculares, pero carecen de una comprensión profunda de cómo se han generado computacionalmente los resultados que ofrecen.

¿Pueden los sistemas actuales de IA generativa, basados en el procesamiento del lenguaje natural o imágenes digitalizadas, sin control humano sobre la interacción con el mundo externo, alcanzar el mismo nivel de razonamiento que los humanos? Por ahora, no parece ser así, pero la IA tampoco es un "loro estocástico", como afirman algunos profesionales. Los científicos de IA no pueden explicar todas las complejidades y propiedades emergentes de los razonamientos inesperados que surgen dentro de las redes neuronales artificiales complejas (12), lo que crea un nuevo y sustancial campo de investigación. La IA es una disciplina científica y tecnológica muy compleja y no sólo —como se dice habitualmente— una herramienta capaz de realizar complejas tareas inteligentes. En medicina, estas complejidades son aún mayores.

PROCESOS CONSCIENTES E INCONSCIENTES E IA

En los primeros sistemas de IA, "prácticos", los llamados sistemas expertos, el objetivo era adquirir el conocimiento y los métodos de razonamiento de los expertos en un dominio específico para transferirlos a un ordenador. Para lograrlo, otra persona, llamada ingeniero del conocimiento, debía extraer ese conocimiento de los expertos mediante métodos como entrevistas, análisis de protocolos y otras técnicas (13) y producir una representación informática. Estos sistemas tuvieron éxito académico y algunas pocas aplicaciones llegaron a ser comerciales durante esos primeros veinte años, pero pocos se llegaron a utilizar con éxito y de forma rutinaria en la práctica clínica, algo que aún ocurre hoy en día con muchos sistemas de IA basados en aprendizaje automático.

¿Por qué no tuvieron éxito estos sistemas expertos en la práctica clínica? Una poderosa razón cognitiva subyacente es que es imposible para un experto verbalizar completamente el conocimiento y los métodos de razonamiento que utiliza, y por tanto, transferirlo a un ordenador. La adquisición de habilidades cognitivas ocurre en varias etapas (14,15), y después de un período de aproximadamente diez años, el razonamiento de un experto -y los procesos cognitivos en general- en un dominio específico se vuelve automatizado y parcialmente inconsciente. En esta etapa posterior, no es posible verbalizar completamente el razonamiento que una persona hace. Esta limitación ha sido un problema central para el "cuello de botella en la adquisición de conocimiento" en los sistemas expertos, convirtiéndose en una barrera para el desarrollo de sistemas expertos generalizables (16).

Los autores recuerdan a un colega, médico experimentado, que comentaba que a veces sentía una sensación difícil de definir, a la que llamaba "gut feeling" - que podríamos traducir por "presentimiento"-, cuando pensaba que algo faltaba en el diagnóstico de un paciente específico, pero no podía comprender ni articular qué era hasta evaluarlo en profundidad. Este tipo de intuiciones, a menudo denominadas bajo el amplio concepto de "ojo clínico", corresponden a razonamientos cognitivos sobre la situación clínica del paciente, a menudo inconscientes, que utilizan los médicos al examinar a un paciente e implica una serie de cuestiones -éticas, sociales, psicológicas, empáticas, de comprensión personal y de experiencias, entre otras- que, por ahora, son muy difíciles de captar para los sistemas pasados y actuales de IA.

INVIERNOS DE LA IA

Las redes neuronales de McCulloch y Pitts (17), junto con la teoría de la información de Shannon y la cibernética de Wiener, impulsaron un campo emergente de reconocimiento de patrones automáticos para modelar la percepción durante la década de 1950. En 1958, el reconocimiento de patrones alcanzó un hito con el Perceptrón, inspirado en la retina (18). Los paralelismos entre este modelo heurístico y los métodos de inferencia estadística se reconocieron rápidamente, aunque en 1968 Minsky publicó el libro titulado "Perceptrones" (19), donde señalaba las graves limitaciones de los modelos matemáticos lineales simples, lo que, junto con los fallos de la traducción automática de idiomas, contribuyó a desencadenar el primer "Invierno de la IA".

Tras vivir personalmente, los dos autores, los períodos en los que se produjeron estos llamados inviernos de la IA, podemos negar la idea de que la IA realmente desapareció o quedó completamente eclipsada durante estos intervalos, como se suele comentar. Cuando se publicó "Perceptrones", el libro de Minsky, la financiación se desplomó durante unos 20 años en todo el campo de las redes neuronales artificiales; pero durante ese tiempo, los sistemas basados en el conocimiento, con una financiación sustancial, florecieron. En cambio, en la década de 1990, cuando los sistemas basados en el conocimiento encontraron dificultades, las redes neuronales artificiales regresaron con éxito, impulsadas por avances como el algoritmo de retropropagación (20), y llegó el "invierno" de los sistemas basados en el conocimiento.

¿Habrá un tercer invierno de la IA, como algunos sugieren? Es muy probable que, entre las numerosas empresas de IA que se están creando, sólo unas pocas sobrevivan; sin embargo, una posible burbuja económica no significaría que la IA vaya a desaparecer o a enfrentarse a una crisis masiva. Determinadas empresas y campos de la investigación y aplicación biomédica seguirán prosperando, mientras que otros desaparecerán.

DATOS, INFORMACIÓN, CONOCIMIENTO... Y TEORÍAS

En las décadas de 1920 y 1930, los pioneros de la física cuántica estaban obsesionados con la acumulación de datos para el avance de las teorías de la disciplina. Esta idea surgió de una sugerencia de Einstein; muchos años después, cuando Heisenberg visitó Princeton, le recordó a Einstein que muchos físicos teóricos se habían centrado en la recopilación y el análisis de datos, siguiendo su sugerencia; sin embargo, durante su diálogo, Einstein admitió su error, tras comprender posteriormente a su idea inicial que las teorías deben guiar la investigación científica, ya que los datos no pueden contribuir, por sí solos, al avance de la ciencia (21).

En la primera mitad del siglo XX, los científicos solían analizar los datos realizando correcciones que frecuentemente favorecían las teorías existentes. Un ejemplo significativo es el de Eddington y su equipo, quienes fueron a la Isla Príncipe en 1919 para verificar, durante un eclipse, las predicciones de Einstein en su teoría de la relatividad, y cometieron errores que se acercaban más a la teoría de Einstein. Eddington estaba convencido de la validez de la teoría de la relatividad, y ese posible sesgo condujo a la confirmación, quizás prematura, de la teoría de la relatividad (22).

Los propios autores experimentaron en un proyecto de aprendizaje automático (1994-97) para extraer reglas de predicción clínica de mil historias clínicas en papel de pacientes con artritis reumatoide, que este conjunto presentaba numerosos problemas. Tras eliminar los casos que tenían numerosos errores, sólo se seleccionaron 340 casos (23). Entre los problemas más importantes de las historias clínicas electrónicas (HCE) actuales se incluyen la recopilación inexacta de datos; interpretaciones equivocadas de médicos y pacientes; cambios en tratamientos a lo largo del tiempo; errores de transcripción; suposiciones y modelos de análisis inadecuados y, especialmente, diferentes tipos de sesgos: selección, clasificación, medición, demográficos, temporales, disponibilidad de datos, algorítmicos, de publicación, etc. Algunos sistemas de IA actuales se han desarrollado utilizando millones de historias clínicas (24), pero muchos de los problemas mencionados pueden estar ocultos en estas bases de datos retrospectivas, cuyos datos se han registrado en diferentes momentos, en diferentes contextos y circunstancias. Tanto la cantidad como la calidad de los datos médicos son clave para garantizar que los proyectos de aprendizaje automático produzcan resultados clínicamente válidos. Todos los desarrolladores y usuarios de estos sistemas deben recordar esta característica fundamental, que lamentablemente a menudo se ignora.

En 2024 se otorgaron los premios Nobel de Física y Química, ambos relacionados con la IA. El Premio Nobel de Química fue otorgado —junto con Baker— a Hassabis y Jumper por el desarrollo de los algoritmos AlphaFold (25), que han permitido predecir la estructura tridimensional de cientos de miles de proteínas. Este logro fue principalmente estadístico, mediante un análisis masivo basado en IA. No obstante, la teoría que subyace al plegamiento de proteínas sigue siendo desconocida y aún no se ha descubierto mediante el uso de algoritmos de aprendizaje automático como AlphaFold y otros. Descubrir teorías científicas sigue siendo más complejo que analizar datos y extraer patrones estadísticos, incluso tan complejos.

EVALUACIONES DE IA

Cuando el sistema Watson de IBM, basado en IA, ganó el concurso televisivo estadounidense Jeopardy! en 2011, IBM decidió ampliar Watson

para abordar aplicaciones médicas, en particular en oncología. Tras un considerable impacto mediático el sistema no alcanzó los resultados esperados. En las numerosas presentaciones públicas realizadas por profesionales de IBM se presentó una evaluación limitada a pocos casos en hospitales concretos -y los autores comentaron este hecho a los creadores en una de estas presentaciones-. Esto suponía un inconveniente significativo, similar al experimentado con muchas de las primeras aplicaciones de IA en medicina. Éstas tendían a tener un rendimiento inferior en hospitales y universidades fuera de sus entornos de desarrollo primario, debido a las diferentes características del conocimiento y los datos disponibles en cada centro (diversidad de datos, sesgos inherentes, procedimientos y protocolos utilizados por los médicos, diferentes tecnologías, etc.).

Un ejemplo es PERFEX, sistema experto para analizar imágenes SPECT cardíacas desarrollado en la década de 1990 en la Universidad de Emory y Georgia Tech, y comercializado con éxito por General Electric, que requirió dos años de diseño e implementación y unos cinco años de evaluación multicéntrica antes de su aprobación para uso clínico (26). Estas evaluaciones multicéntricas han demostrado ser esenciales para garantizar la validez de los proyectos de IA en medicina. En los últimos años, miles de sistemas de IA han sido completados para uso clínico, pero muchos aún necesitan someterse a evaluaciones sistemáticas en la práctica clínica habitual y en centros distintos a los que los desarrollaron, lo que incluye el cumplimiento de los requisitos exigidos por agencias públicas. Podría ser necesario esperar varios años para determinar su verdadero impacto en la práctica clínica.

LA IA COMO EL ORÁCULO DE DELFOS, UN SUSTITUTO PARA LOS MÉDICOS

Uno de los primeros y más conocidos sistemas de IA en medicina fue INTERNIST-I, un sistema experto cuyo objetivo era gestionar el conocimiento de numerosas enfermedades, similar al amplio ámbito de práctica de un médico internista. Sus creadores afirmaron que el estilo de consulta diagnóstica del programa INTERNIST-I se asemejaba a un "oráculo griego" (27), pero finalmente terminó transformándose en un sistema de referencia llamado Referencia Médica Rápida (QMR), en lugar de un verdadero oráculo médico.

En los últimos años, han surgido sistemas de IA generativa, capaces de responder a numerosas cuestiones médicas (diagnósticos, tratamientos, prevención, etc). El concepto de oráculo no era útil en los sistemas pioneros de la IA, y nos encontramos ante una posible similitud hoy con los sistemas de IA generativa. Algunos sistemas actuales de IA, cuyos autores afirman ser capaces de predecir resultados para cientos de enfermedades -en uno de ellos, llamado DELFOS 2-M (28), no parece casualidad el nombre elegido por sus creadores-, aún deben demostrar dichos resultados con evaluaciones sistemáticas.

Los defensores de los sistemas generativos de IA sostienen que, con su análisis de grandes cantidades de datos, contienen en su interior el tipo de razonamiento y enfoques lógicos utilizados por expertos humanos; pero estos procesos humanos incluyen procesos como la intuición, el sentido común, las emociones, la empatía, las consideraciones éticas, las generalizaciones a nuevos casos, la comprensión de problemas psicosociales pasados y del entorno, etc., aún muy diferentes a la IA.

EXPLICABILIDAD E INTERPRETABILIDAD

Un componente clave de la IA, aún en desarrollo y sin resolver, es la llamada IA explicable, cuyo objetivo es comprender y describir claramente cómo y por qué un sistema de IA toma una decisión o predicción específica.

Un problema fundamental en la IA médica en los últimos años es cómo lograr que un sistema de IA explique las razones detrás de una decisión o resultado; por ejemplo un diagnóstico o una recomendación terapéutica específica, y cómo garantizar que las conclusiones del sistema sean fácilmente interpretables por el usuario. Los autores han publicado recientemente una revisión sobre IA explicable en medicina (29).

Algunas —o muchas— personas que trabajan en IA, sin muchos años de experiencia en el campo, piensan que la explicabilidad de los sistemas de IA es un tema reciente, olvidando que se ha estudiado durante más de cincuenta años. El concepto de "caja negra" —un sistema que no permite explicar los detalles de sus procesos internos— es común en los sistemas basados en redes neuronales, pero la explicabilidad era ya esencial en el diseño de los sistemas basados en el conocimiento. De hecho, dos de los primeros sistemas expertos médicos, MYCIN (30) y CASNET (31,32), incluían explicaciones de sus resultados. MYCIN contaba con un módulo de explicación que mostraba al usuario la lista de reglas activadas y utilizadas para alcanzar el resultado final del sistema. CASNET, creado por uno de los autores, proporcionaba al usuario una explicación del fundamento causal o asociativo de sus conclusiones.

El siguiente gráfico muestra los diferentes tipos de explicabilidad/interpretabilidad. Podemos observar que los sistemas basados en reglas tradicionales SI…ENTONCES… (típicas de los sistemas expertos) presentan una menor exactitud, aunque con mayor capacidad explicativa e interpretativa, mientras que en los sistemas de aprendizaje profundo ocurre lo contrario.

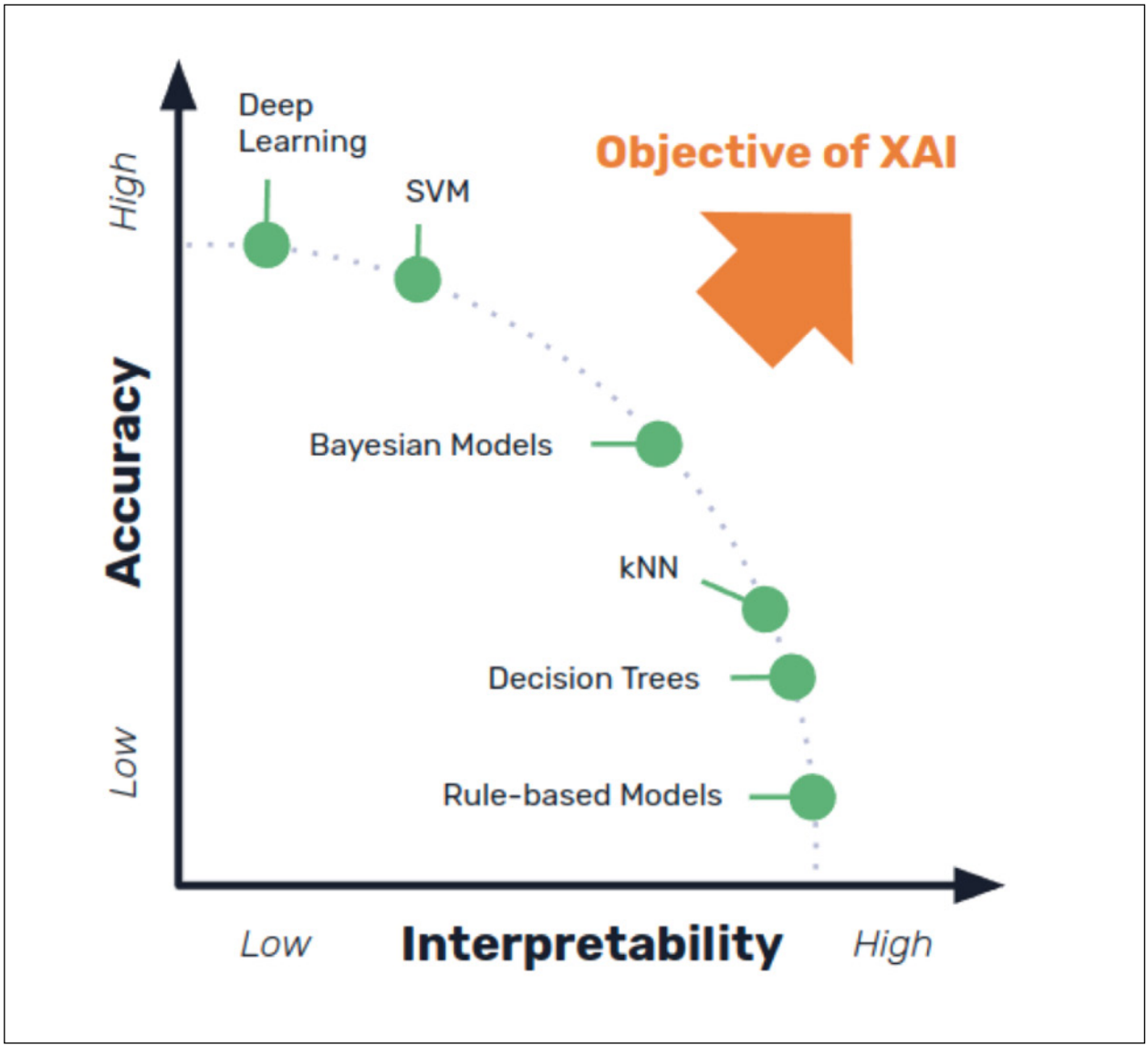

Figura 1. Relación entre exactitud e interpretabilidad para diferentes técnicas de IA. Publicado originalmente por los autores en (29).

Las conclusiones de los sistemas expertos se podían explicar gracias a que su razonamiento y conocimiento eran explícitos y se conocían en detalle, mientras que los modelos generativos, basados en sus capacidades predictivas, al ser redes neuronales artificiales aún se comportan como cajas negras. La explicabilidad de los modelos grandes de lenguaje es limitada, aunque se han propuesto técnicas como el llamado "chain-of-thought", el uso de otros LLMs o modelos híbridos, neurosimbólicos.

Un estudio reciente mostró que un amplio porcentaje de los médicos que habían utilizado un sistema de IA generativo consideraba que las explicaciones correctas del sistema no eran un problema particularmente relevante, siempre que el sistema proporcionara buenas decisiones (33). Podríamos decir que, si los sistemas pioneros de la IA se encontraron con una desconfianza abrumadora de los médicos clínicos, en muchos casos encontramos ahora la situación contraria.

RAZONAMIENTO MÉDICO E IA

Las estrategias de razonamiento para la resolución de problemas utilizadas por los médicos en la práctica clínica incluyen diferentes estrategias, como el uso de heurísticas, la aplicación de diferentes tipos de lógica (deductiva, inductiva y abductiva), la estimación de probabilidades de enfermedades basándose en comparaciones entre ellas en lugar de estadísticas, el reconocimiento rápido de patrones textuales y visuales, el razonamiento causal, la experiencia con casos previos o el sentido común, entre otros.

Varios sistemas de IA desarrollados en los primeros treinta años de aplicaciones médicas (sistemas basados en conocimientos) podían abordar con cierto éxito cada uno de estos problemas por separado. Por ejemplo, los sistemas IA pioneros para diagnóstico incluían capacidades de razonamiento hipotético-deductivo; los sistemas expertos como CASNET o ABEL incluían razonamiento causal que relacionaba los síntomas, signos y pruebas utilizadas con los procesos fisiopatológicos subyacentes de una enfermedad específica; y los sistemas de razonamiento basado en casos podían comparar el caso de un nuevo paciente con una biblioteca de casos previamente almacenada, seleccionando los más similares y proponiendo una estrategia terapéutica. La diferencia con el razonamiento utilizado por los profesionales médicos es evidente, ya que los médicos son capaces de integrar todas las tareas mencionadas (y otras), mientras que estos sistemas sólo podían simular una de ellas.

La IA generativa busca unificar todas estas estrategias, extrayendo conocimiento y estrategias de razonamiento de grandes cantidades de datos, incluyendo artículos científicos, historiales médicos, registros clínicos, libros, sitios web, etc. Los científicos que desarrollan IA generativa argumentan que este conocimiento y estas estrategias de razonamiento están, en última instancia, integrados en los datos científicos procesados por los sistemas, que son capaces de realizar internamente inferencias y asociaciones complejas que conducen a resultados similares a los de los médicos. Por el contrario, carecen de experiencia clínica real (como la de los médicos, que pueden valorar el contexto psicológico y social completo de sus pacientes), no pueden comprender las relaciones causales y los procesos fisiopatológicos que ya estaban representados en sistemas expertos como CASNET o ABEL, ni tienen la capacidad de comprender todos los aspectos éticos de la medicina, que los médicos aprenden a lo largo de años de experiencia.

Los modelos grandes de lenguaje (LLMs, acrónimo en inglés) como ChatGPT, Claude, y otros, generalmente pueden producir simulaciones muy coherentes de razonamiento deductivo, pero carecen de conocimiento intrínseco de sus fundamentos lógicos, ya que extraen este conocimiento de las estadísticas de patrones lingüísticos de millones de textos. Por lo tanto, pueden cometer errores en las deducciones porque no pueden verificar internamente la verdad de las premisas o no tienen información correcta suficiente y pueden ser propensos a proponer respuestas inconsistentes o falacias como las "alucinaciones", conclusiones que parecen correctas pero que son lógicamente erróneas. De nuevo, aunque algunos de estos LLMs pueden superar claramente a los profesionales humanos en diversas tareas —en particular, cuando se deben procesar cálculos complejos y grandes cantidades de datos o conocimientos—, estos métodos de razonamiento de los LLMs no pueden sustituir las múltiples capacidades, habilidades y la experiencia juzgada de los mejores médicos.

LIMITACIONES Y RIESGOS DE LA IA

Pioneros de los sistemas expertos y de ayuda a las decisiones clínicas, como Feigenbaum, Buchanan, Kulikowski, Shortliffe, Szolovits, Barnett, Greenes y otros, ya comprendían las numerosas limitaciones de la IA y los peligros de aceptar sus decisiones sin una comprensión profunda de su funcionamiento y sin las explicaciones válidas generadas por estos sistemas. Ya en 2004 los autores advertían sobre estas limitaciones (34). A continuación, destacamos algunas limitaciones críticas de la IA actual.

SENTIDO COMÚN

Un dicho popular (internacional) dice que "el sentido común es el menos común de los sentidos". Aristóteles lo consideraba la capacidad de las personas para formarse juicios coherentes sobre el mundo.

Turing afirmó que la lógica matemática no puede ser suficiente para sustentar la razón sin considerar el sentido común (35). Fue el primero en el mundo científico relacionado con los ordenadores y la IA en mencionar explícitamente el sentido común. Valiant (10) menciona que nos enfrentamos a dos problemas relacionados con el sentido común: identificar qué es lo que la lógica no logra captar —una consecuencia de que la lógica matemática requiere un marco teórico sólido para funcionar correctamente— y determinar la vía científica necesaria para abordar el problema del sentido común —para lo cual necesitamos, paradójicamente, una teoría general de lo no teórico—. En los albores de la IA simbólica en la década de 1950, investigadores como John McCarthy (36) y Marvin Minsky (37) identificaron el «problema del sentido común»: la dificultad de dotar a las máquinas de conocimientos básicos sobre el mundo.

Los humanos son particularmente hábiles para generalizar, incluso a partir de unos pocos ejemplos, pero en la era de los sistemas basados en el conocimiento, las máquinas no podían razonar más allá del conjunto de reglas proporcionadas por sus diseñadores, por lo que a menudo cometían errores en situaciones nuevas o implícitas. Por otro lado, los modelos generativos no utilizan reglas explícitas; en cambio, aprenden patrones estadísticos a partir de cantidades masivas de datos (textos o imágenes). Los desarrolladores de IA generativa argumentan que estos sistemas capturan el conocimiento de sentido común implícito en los textos con los que se entrenan o en las inferencias que pueden extraer; sin embargo, si bien se propone que la IA generativa podría mejorar este enfoque, aún no puede considerarse equivalente a la de los humanos, ya que estos modelos no razonan realmente ni tienen una comprensión propia del mundo físico.

ALUCINACIONES

Las alucinaciones en modelos grandes de lenguaje como ChatGPT, Claude, y muchos otros, ocurren cuando estos modelos generan información falsa o inexacta, pero lo hacen de una manera que puede parecer plausible y coherente a los usuarios. Las causas pueden ser diversas (38), como, por ejemplo, limitaciones en los datos de entrenamiento, errores, sesgos, información incompleta o la invención de información falsa, una falta de comprensión del contexto, la formulación inadecuada de la pregunta planteada al sistema, o cuando el sistema no encuentra una respuesta y responde incorrectamente. Es importante recordar que estos sistemas no comprenden realmente las respuestas que proporcionan. Por lo tanto, pueden proporcionar respuestas que, desde un punto de vista lingüístico, parecen correctas, pero que, en realidad, contienen información falsa.

Entre la inmensa cantidad de datos utilizados para entrenar estos sistemas —que algún día podrían abarcar todos los documentos creados por personas, como una biblioteca universal Borgiana—, con frecuencia existen numerosos errores, sesgos y lagunas en el conocimiento científico sobre muchos temas. Los LLMs pueden proporcionar respuestas que a menudo parecen plausibles, pero en realidad son fallos en la generación o el razonamiento del modelo, debidos a problemas como limitaciones en los datos de entrenamiento —en calidad o cantidad, por ejemplo, o en la falta de conocimientos ampliamente aceptados sobre el tema—, entre otras causas. De esta forma las alucinaciones pueden llevar a médicos y pacientes a malas decisiones clínicas, que se pueden sumar a otras causas de posible mala praxis por el uso de la IA en la práctica clínica, con responsabilidad compartida por el profesional sanitario (39).

Una necesidad evidente en la práctica clínica es comprender cómo se producen las alucinaciones y sus causas, algo que un médico debe conocer para poder utilizar un sistema de IA conociendo sus limitaciones. Estos errores son inherentes a cualquier sistema basado en datos y análisis estadísticos.

INCERTIDUMBRE Y RIESGO

"Incertidumbre" es una de las palabras que mejor resume las numerosas dificultades a las que se enfrentan los investigadores de IA en biomedicina. La incertidumbre se encuentra en numerosos aspectos de la documentación o el razonamiento médico, como los múltiples errores e inconsistencias que encontramos en los datos de las historias clínicas electrónicas; las declaraciones subjetivas de los pacientes y registradas en sus historiales; el razonamiento implícito integrado en los datos recopilados, correspondiente a decisiones tomadas por los médicos pero no registradas en estos documentos; la falta de conocimiento, aún por descubrir, en tantas áreas de la medicina, incluyendo las causas de las enfermedades; las diferencias en los protocolos y la gestión de pacientes entre diferentes profesionales en distintas consultas médicas; los errores y discrepancias entre los dispositivos médicos utilizados en cada hospital o clínica, etc. Todos estos factores contribuyen al riesgo de las decisiones y propuestas terapias médicas, para lo cual las teorías estadísticas actuales basadas aún en la economía utilitaria resultan completamente inadecuadas para tomar en cuenta las consideraciones particulares del paciente individual y su tratamiento ético sujeto al juicio Hipocrático.

Los modelos de IA generativa son esencialmente probabilísticos. No tienen forma de saber cuándo tienen incertidumbre sobre una respuesta, a menudo pueden inducir a error a los usuarios, y representan implícitamente la incertidumbre mediante distribuciones de probabilidad en el conjunto de datos que manejan. Para superar este problema clave, los desarrolladores de modelos de IA generativa han creado diferentes métodos, pero estos todavía no incluyen nuevas representaciones del sentido común y contexto humano que se necesita para la práctica ética de la medicina.

CONCLUSIONES

La adquisición acelerada de nuevas habilidades cognitivas mediante la IA también puede llevar, por el contrario, a la pérdida de otras habilidades tradicionales (40). Por ejemplo, artículos recientes (41,42). advierten sobre la pérdida de habilidades —p. ej., razonamiento avanzado en casos médicos complejos—. Tras pocos años desde la aparición de ChatGPT han aparecido numerosos "expertos" en IA, sin experiencia real, y, lo que es peor, miles de artículos científicos sobre el uso de la IA en medicina, cada año, con diseños de estudio de validez más que dudosa en muchos casos.

Existe un probable exceso de expectativas actuales sobre la IA y sus capacidades, especialmente en medicina, pero también existe una gran esperanza de que la IA se convierta en el centro de una revolución científica comparable a otras anteriores en medicina. Numerosos artículos han sugerido, incluso desde los años 1970 (43), que sistemas de IA superan en resultados a los médicos —con un claro ejemplo, el diagnóstico de mamografías en estudios recientes (44)—, aunque al final el médico aún debe supervisar los resultados de la IA para evitar posibles consecuencias negativas en los pacientes. En países africanos, por ejemplo, el uso de la IA podrá suponer un cambio sustancial en la asistencia sanitaria futura, debido a la falta de médicos.

En este artículo hemos seleccionado diez lecciones aprendidas, pero podrían ser más. Por ejemplo, en el momento de concluir este artículo se ha publicado un amplio experimento mostrando los

problemas del uso de un estetoscopio basado en IA (45), con resultados mejores que los médicos con los que se comparó, pero que no es eficiente debido al tiempo necesario para su uso y su falta de integración clínica —algo que ya se había visto con los primeros sistemas de IA médica en los años 1970s (46)—. En este contexto, los profesionales médicos necesitan conocer los fundamentos básicos y limitaciones de estos sistemas de IA así como las lecciones aprendidas tras más de cincuenta años de experiencias.

AGRADECIMIENTOS

Víctor Maojo tiene apoyo de proyectos propios de la Universidad Politécnica de Madrid y del proyecto europeo SHIELD (European Union's Horizon Europe research and innovation program under grant agreement NA 101156751.

DECLARACIÓN DE TRANSPARENCIA

Los autores/as de este artículo declaran no tener ningún tipo de conflicto de intereses respecto a lo expuesto en el presente trabajo.

BIBLIOGRAFÍA

1. Ledley, RS and Lusted, LB. Reasoning foundations of medical diagnosos: symbolic, logic, probability and value theory aid our understanding of how physicias reason. Science 1959: 130 (3366): 9-21.
2. Pauker SG, Gorry GA, Kassirer JP et al. Towards the simulation of clinical cognition. Taking a present illness by computer. Am J Med. 1976 Jun;60(7):981-96. doi: 10.1016/0002-9343(76)90570-2. PMID: 779466.
3. Gorry A. Strategies for computer-aided diagnosis. Math Biosci 1968; 2:293-318.
4. Warner, HR, Toronto, AF, Veasey, LG. Et al. A mathematical approach to medical diagnosis: application to congenital heart disease. JAMA 1961: 177 (3): 177-83
5. Tversky, A. and Kahneman, D. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.Science. 1974 Vol 185, Issue 4157. pp. 1124-1131
6. Feinstein AR. Clinical Judgment. Baltimore: Williams and Wilkins; 1967.
7. Simon HA, CA Kaplan. C.A. in Foundations of Cognitive Science. M. Posner (Ed): MIT Press, Cambridge, MA; 1990.
8. Wachter, R. The Digital Doctor: Hope, Hype and Harm at the Dawn of Medicine's Computer Age. McGraw Hill. 2015.
9. McCarthy, J., Minsky, M. L., Rochester, N., el al. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Unpublished report. 1956
10. Valiant, L.; Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World. Basic Books. 2014
11. Jelinek, F. Statistical Methods for Speech Recognition. The MIT Press, 1997
12. Amodei, D. The Urgency of Interpretability. https://www.darioamodei.com/post/the-urgency-of-interpretability
13. Pazos, J. Inteligencia Artificial. Paraninfo.1987
14. Musen MA, van der Lei J. Knowledge engineering for clinical consultation programs: modeling the application area. Methods Inf Med. 1989 Jan;28(1):28-35.
15. Maojo, V. Cerebro y Música. Entre la neurociencia, la tecnología y el arte. EMSE EDAPP. 2018
16. Gomez, A.; Juristo, N; Montes, C. et al. Ingenieria del Conocimiento. Editorial Centro de Estudios Ramón Areces. Madrid, Spain. 1997
17. McCulloch, W. and Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity". The Bulletin of Mathematical Biophysics. 1943 Vol. 5, pp. 115-133
18. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review. 1958 Vol. 65, No. 6, pp. 386–408
19. Minsky, M. and Papert, S.A. Perceptrons: An Introduction to Computational Geometry. MIT Press, Cambridge.1969
20. Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. Nature 1986 323, 533–536 . https://doi.org/10.1038/323533a0
21. Heisenberg Encuentros y conversaciones con Einstein y otros ensayos . Alianza Editorial., 1979
22. Dyson, F.W.; Eddington, A.S.; Davidson, C.R. (1920) "A determination of the deflection of light by the sun's gravitational field, from observations made at the solar eclipse of May 29, 1919," Philosophical Transactions of the Royal Society A 220: 571-581.
23. Sanandres, Ja.; Maojo, V.; Crespo, J et al.. A clustering-based constructive induction method and its application to Rheumatoid arthritis. Proceedings of AI in Medicine. 2101. 2001 pp. 59 - 62.
24. Callaway E. Medical AI trained on whopping 57 million health records. Nature. 2025 May 6. doi: 10.1038/d41586-025-01422-3
25. Jumper J, Evans R, Pritzel et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583-589. doi: 10.1038/s41586-021-03819-2.
26. Garcia EV, Cooke CD, Folks RD et al. Diagnostic performance of an expert system for the interpretation of myocardial perfusion SPECT studies. J Nucl Med. 2001 Aug;42(8):1185-91.
27. Miller RA, Masarie FE Jr. The demise of the "Greek Oracle" model for medical diagnostic systems. Methods Inf Med. 1990 Jan;29(1):1-2.

28. Shmatko A, Jung AW, Gaurav K, et al. Learning the natural history of human disease with generative transformers. Nature. 2025 Nov;647(8088):248-256. doi: 10.1038/s41586-025-09529-3.
29. González-Alday, R., García-Cuesta, E., Kulikowski, C. A., et al. A Scoping Review on the Progress, Applicability, and Future of Explainable Artificial Intelligence in Medicine. Applied Sciences 2023, 13(19), 10778.
30. Shortliffe EH. Computer Based Medical Consultations: MYCIN New York: Elsevier; 1976.
31. Kulikowski CA, Weiss SM. Representation of expert knowledge for consultation: the CASNET and EXPERT projects. In: Szolovits P, editor. Artificial Intelligence in Medicine. AAAS Selected Symposia Series. Boulder CO: Westview Press; 1982:21-55.
32. Kulikowski, C.A. Strategies for test selection in causal network models.Tech. Report #TR-11. Computers in Biomedicine. Rutgers University. 1972
33. Sumner J, Wang Y, Tan SY, Chew EHH et al. Perspectives and Experiences With Large Language Models in Health Care: Survey Study. J Med Internet Res. 2025 May 1;27:e67383. doi: 10.2196/67383.
34. Maojo, V. Domain-Specific Particularities of Data Mining: Lessons Learned. Proceedings of ISBMDA (2004). Lecture Notes in Computer Science. Pp 235-242.
35. Turing, A. M. Solvable and Unsolvable Problems. Science News, **31**, 7–23. Penguin Books, Melbourne-London-Baltimore. 1954
36. McCarthy, J. Circumscription—A form of non-monotonic reasoning. Artificial Intelligence 1980, 13(1-2), 27–39
37. Minsky, M. The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon & Schuster. 2006.
38. Yubin K., Hyewon J., Shan Chen, Sh. et al. Medical Hallucination in Foundation Models and Their Impact on Healthcare. medRxiv 2025.02.28.25323115
39. Wu D, Haredasht FN, Maharaj SK,et al. First, do NOHARM: towards clinically safe large language models. ArXiv [Preprint]. 2025 Dec 17:arXiv:2512.01241v2.
40. Bastani H, Bastani O, Sungu A et al. Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proc Natl Acad Sci U S A. 2025 Jul;122(26):e2422633122. doi: 10.1073/pnas.2422633122.
41. Abdulnour RE, Gin B, Boscardin CK. Educational Strategies for Clinical Supervision of Artificial Intelligence Use. N Engl J Med. 2025 Aug 21;393(8):786-797. doi: 10.1056/NEJMra2503232.
42. Fogo AB, Kronbichler A, Bajema IM. AI's Threat to the Medical Profession. JAMA. 2024;331(6):471–472. doi:10.1001/jama.2024.0018
43. Yu VL, Fagan LM, Wraith SM et al. Antimicrobial selection by a computer. A blinded evaluation by infectious diseases experts. JAMA. 1979 Sep 21;242(12):1279-82.
44. Gommers J, Hernström V, Josefsson et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. Lancet. 2026 Jan 31;407(10527):505-514. doi: 10.1016/S0140-6736(25)02464-X.
45. Kelshiker MA, Bächtiger P, Petri CF et al. Triple cardiovascular disease detection with an artificial intelligence-enabled stethoscope (TRICORDER) in the UK: a cluster-randomised controlled implementation trial. Lancet 2026 Feb 14;407(10529):704-715. doi: 10.1016/S0140-6736(25)02156-7.
46. Buchanan, B.G. and Shortliffe, E.H. Rule-based expert systems: The Mycin experiments of the Stanford heuristic programming Project. Addison-Wesley, Reading, MA.1984

Si desea citar nuestro artículo:
Maojo V, Kulikowski CA. Inteligencia Artificial y medicina: diez lecciones aprendidas (y olvidadas): 1970-2026. An RANM. 2026;143(01): 67–75. DOI: 10.32440/ar.2026.143.01.rev05