IA y sanidad: algunas consideraciones sobre la confianza

Marcos Alonso. Universidad Complutense de Madrid. España
marcos.alonso@ucm.es
Ramón Ortega Lozano. Escuela de Enfermería y Fisioterapia San Juan de Dios. Universidad Pontificia Comillas. España
rortegal@comillas.edu
Aníbal M. Astobiza. Universidad de Granada. España
anibalmastobiza@gmail.com
Recibido 28/06/2025 • Aceptado 30/10/2025

Resumen

En este artículo exploramos la naturaleza relacional de la confianza en los sistemas de inteligencia artificial aplicados al ámbito médico. Frente a los enfoques que tienden a localizar la confianza en propiedades individuales —ya sea del sujeto que confía o del objeto en quien se confía—, proponemos comprenderla como un fenómeno emergente que solo puede desplegarse plenamente en la interacción entre humanos y tecnologías. Analizamos críticamente algunos de los conceptos clave del debate actual —como la responsabilidad, la rendición de cuentas, el antropomorfismo y la alineación de valores—, mostrando que todos ellos resisten ser atribuidos unívocamente a uno de los polos de la relación. Sostenemos que una comprensión adecuada de estos constructos exige situarlos en una perspectiva relacional, donde la confianza no se deriva simplemente de cualidades técnicas o actitudes subjetivas, sino de estructuras compartidas de sentido, prácticas de corresponsabilidad y marcos institucionales adecuados. Este enfoque permite afrontar con mayor precisión los desafíos éticos de la IA médica y orienta el diseño de sistemas que no solo sean eficientes, sino también dignos de confianza en un sentido robusto.

Palabras clave: Inteligencia artificial, confianza, ética aplicada.

Abstract

This article explores the relational nature of trust in artificial intelligence (AI) systems applied to the medical field. In contrast to approaches that tend to locate trust in individual properties—either of the trusting subject or of the object of trust—we propose understanding it as an emergent phenomenon that can only unfold fully within the interaction between humans and technologies. We critically analyze several key concepts in the current debate—such as responsibility, accountability, anthropomorphism, and value alignment—showing that none of them can be unequivocally attributed to a single pole of the relationship. We argue that an adequate understanding of these constructs requires situating them within a relational perspective, where trust does not simply derive from technical qualities or subjective attitudes, but from shared structures of meaning, practices of co-responsibility, and appropriate institutional frameworks. This approach allows for a more precise engagement with the ethical challenges of medical AI and guides the design of systems

ISSN-e: 1885-5679

that are not only efficient but also trustworthy in a robust sense.

Keywords: artificial intelligence, trust, applied ethics

ISSN-e: 1885-5679

IA y sanidad: algunas consideraciones sobre la confianza

Marcos Alonso. Universidad Complutense de Madrid. España
marcos.alonso@ucm.es
Ramón Ortega Lozano. Escuela de Enfermería y Fisioterapia San Juan de Dios. Universidad Pontificia Comillas. España
rortegal@comillas.edu
Aníbal M. Astobiza. Universidad de Granada. España
anibalmastobiza@gmail.com
Recibido 28/06/2025 • Aceptado 30/10/2025

1. Introducción: Planteamiento del problema y temas a desarrollar.

La creciente incorporación de sistemas de inteligencia artificial (IA) en la práctica médica ha reavivado con fuerza una vieja preocupación filosófica: la cuestión de la confianza. En contextos clínicos, donde la incertidumbre, la vulnerabilidad y la asimetría de información son inherentes, confiar en otros —especialmente en quienes toman decisiones sobre nuestra salud— ha sido tradicionalmente una necesidad ética y epistémica. Hoy, sin embargo, esa necesidad parece proyectarse también sobre agentes no humanos. ¿Podemos —o debemos— confiar en sistemas de IA? ¿Qué tipo de confianza está en juego cuando las decisiones médicas comienzan a apoyarse en algoritmos? Y, más fundamentalmente, ¿estamos hablando de lo mismo cuando hablamos de confiar en una persona y confiar en una tecnología?

El uso creciente de IA en medicina agudiza esas tensiones. Tal como advierten Vereschak et al. (2021), los sistemas de IA en salud pueden comprometer la seguridad, la equidad y la dignidad humana. Problemas como los sesgos algorítmicos, la opacidad de los modelos (las llamadas "cajas negras") y la recolección masiva de datos con fines de entrenamiento, exponen a pacientes y profesionales a nuevos riesgos. A esto se suman dificultades normativas para la atribución de responsabilidades cuando las decisiones clínicas se ven mediadas —o condicionadas— por algoritmos. Como señalan Sparrow y Hatherley (2019), la promesa de una medicina más precisa y

ISSN-e: 1885-5679

eficiente puede terminar reforzando dinámicas de vigilancia y exclusión, especialmente si no se problematizan las condiciones sociales y políticas en las que estas tecnologías son diseñadas e implementadas. Frente a este panorama, se ha popularizado en los últimos años el paradigma de la "IA digna de confianza" o *Trustworthy AI* (Floridi, 2019; Thiebes et al., 2020). Esta formulación responde a un anhelo —no exento de ambivalencias— de restaurar cierta estabilidad en un mundo técnico-científico percibido como cada vez más incierto. Como apuntan Choung et al. (2022), la confianza es un mecanismo humano fundamental para afrontar la ambigüedad, la complejidad y la incertidumbre. En este sentido, el énfasis en la *trustworthiness* parece actuar como una respuesta compensatoria: se proyecta sobre la tecnología la esperanza de que será capaz de ofrecernos seguridad allí donde las condiciones sociales y políticas fallan en garantizarla.

Sin embargo, hablar de una "IA confiable" exige afinar el concepto de confianza. Proponemos, en este sentido, distinguir entre tres dimensiones: el sujeto que confía (trustor), el objeto en quien se deposita la confianza (trustee) y el contexto relacional que condiciona dicha interacción. Esta triada permite superar visiones reduccionistas que se centran exclusivamente en las propiedades técnicas de los sistemas, sin considerar las disposiciones del usuario ni el entramado institucional en el que se inserta la tecnología. Desde esta perspectiva, el trustor no es una figura genérica, sino un agente situado, cuyas experiencias, expectativas y propensión a confiar están moldeadas por factores sociales, culturales y afectivos. A su vez, el trustee —en este caso, el sistema de IA— no puede considerarse confiable en el mismo sentido que una persona, dado que carece de intencionalidad, responsabilidad moral o capacidad de respuesta ética. Aquí se impone una distinción clásica pero frecuentemente ignorada: la diferencia entre *trustworthiness* y *dependability*. Mientras que la primera remite a cualidades morales como la honestidad, la integridad o la benevolencia, la segunda alude a la consistencia funcional de un sistema. Esta distinción resulta fundamental: una IA puede ser altamente *dependable* sin ser, en sentido estricto, *trustworthy*.

En suma, la irrupción de la inteligencia artificial en el ámbito médico no exige únicamente ajustes técnicos o normativos. Reclama, sobre todo, una revisión conceptual rigurosa de categorías como confianza, responsabilidad y agencia. Frente a los discursos que promueven la ilusión de una "IA confiable", creemos que una

ISSN-e: 1885-5679

mirada filosófica puede ofrecer herramientas críticas para pensar no tanto en qué —o en quién— confiamos, sino en qué condiciones puede emerger una confianza legítima, informada y ética en el corazón mismo de la práctica clínica.

2. Desarrollo: Encuadre teórico-metodológico y análisis de la información

La inteligencia artificial (IA) ha comenzado a desempeñar un papel cada vez más destacado en la asistencia sanitaria. Diversos estudios han documentado cómo los sistemas basados en IA pueden igualar o incluso superar las capacidades humanas en tareas clínicas fundamentales, como el análisis de historiales médicos, la interpretación de datos clínicos, el diagnóstico por imagen o el diseño de tratamientos personalizados (Topol, 2019). No estamos, pues, ante una herramienta neutra como un bisturí o un estetoscopio, sino ante una tecnología que reconfigura la distribución de autoridad práctica y epistémica entre humanos y máquinas (Choung et al., 2022). La confianza, en este nuevo escenario, adquiere un relieve especial. Tradicionalmente, en el ámbito médico, se ha considerado que la confianza entre médico y paciente tiene un valor intrínseco, al constituir la base de una relación humana significativa, pero también un valor instrumental, al facilitar la adherencia terapéutica y mejorar los resultados clínicos (Hatherley, 2020). Ahora bien, cuando se traslada este concepto a las tecnologías de IA, surgen tensiones fundamentales: ¿podemos confiar en una entidad sin voluntad, sin emociones ni intenciones?

Diversos autores han subrayado que la confianza no se reduce a la fiabilidad: confiar en alguien implica asumir que actúa con buena voluntad y con motivaciones apropiadas (Hatherley, 2020). En este sentido, los sistemas de IA, por carecer de agencia moral, no serían verdaderos candidatos a ser objeto de confianza. Pero esta visión, aunque relevante, no agota la complejidad del problema. En lugar de centrarnos exclusivamente en la falta de intencionalidad de la IA, proponemos una reconstrucción del fenómeno de la confianza en clave relacional, que permita identificar los distintos factores en juego sin presuponer una equivalencia entre confiar en personas y confiar en tecnologías. Para ello, hemos desarrollado una clasificación tripartita que distingue entre dimensión subjetual, dimensión objetual y dimensión relacional de la confianza. Esta propuesta no busca compartimentar el fenómeno, sino ofrecer un marco que permita analizar sus diversas facetas desde una perspectiva

ISSN-e: 1885-5679

fenomenológicamente orientada, que reconozca la primacía de la relación como estructura fundamental.

La dimensión subjetual hace referencia a los factores que dependen del sujeto que confía: sus expectativas, creencias, conocimientos previos y experiencias con tecnologías similares. Preferimos hablar de "subjetual" en lugar de "subjetivo" para evitar connotaciones relativistas y contraposiciones simplistas con lo "objetivo". En el caso de la IA médica, el modo en que los usuarios interpretan y evalúan el comportamiento del sistema está fuertemente influido por su familiaridad con la tecnología.

Un segundo factor subjetual crucial es el conocimiento. La comprensión —aunque sea básica— del funcionamiento de un sistema de IA puede reducir la ansiedad e incrementar la confianza del usuario. Esta dimensión se conecta directamente con el debate sobre la interpretabilidad y la explicabilidad de los sistemas automatizados. Ahora bien, mientras que la comprensión se sitúa del lado del sujeto, la explicabilidad es un atributo del objeto, y por tanto pertenece a la dimensión objetual. La dimensión objetual incluye aquellos factores vinculados con las propiedades intrínsecas del sistema de IA. Entre ellos, destacan la capacidad, la precisión, la fiabilidad y, como ya mencionamos, la explicabilidad. La capacidad hace referencia a la competencia técnica de la IA para ejecutar las tareas para las que fue diseñada (Malle & Ullman, 2021). En un entorno tan sensible como el médico, esta competencia no solo se refiere al rendimiento bajo condiciones ideales, sino también a la adaptabilidad del sistema frente a contextos variables.

La precisión —entendida como la habilidad del sistema para minimizar errores diagnósticos o terapéuticos— tiene un peso evidente en la generación de confianza, especialmente cuando los riesgos asociados a decisiones incorrectas pueden ser graves o irreversibles. De igual modo, la fiabilidad, definida como la consistencia en el funcionamiento del sistema a lo largo del tiempo (London, 2019), es una condición indispensable para que se genere una relación de confianza duradera. La explicabilidad, por su parte, se refiere a la capacidad del sistema para ofrecer justificaciones comprensibles de sus decisiones (Shin, 2021). Aunque no garantiza por sí sola la confianza, su ausencia puede dificultarla, especialmente en contextos donde la opacidad algorítmica impide al usuario comprender por qué se ha tomado una

ISSN-e: 1885-5679

determinada decisión clínica. Finalmente, la dimensión relacional busca integrar las dos anteriores reconociendo que la confianza no es una simple proyección de propiedades individuales (ya sean del sujeto o del objeto), sino un fenómeno emergente que se constituye en la interacción. Esta dimensión relacional es especialmente importante cuando hablamos de tecnologías que, como la IA médica, se introducen en una práctica que ya está profundamente estructurada por relaciones humanas, normativas y simbólicas.

Desde esta perspectiva, la confianza en la IA médica no puede abordarse únicamente como un problema técnico o psicológico, sino que debe ser entendida como una configuración relacional que articula diferentes niveles: la percepción del usuario, las propiedades del sistema y el contexto institucional donde tiene lugar la interacción. Es en este cruce donde se juegan las cuestiones más relevantes: ¿cómo se distribuye la responsabilidad cuando una IA participa en un diagnóstico? ¿Qué tipo de vínculo se establece entre el paciente y un sistema no humano? ¿Cómo se preserva la autoridad clínica sin delegarla ciegamente en algoritmos?

Responder a estas preguntas requiere una aproximación interdisciplinar y crítica, capaz de integrar elementos procedentes de la filosofía, la ética, la sociología y las ciencias computacionales. Pero también exige una cierta modestia epistémica: si bien la IA ofrece oportunidades extraordinarias para mejorar la atención sanitaria, no debemos perder de vista que su introducción transforma las condiciones mismas de la relación clínica. En este nuevo escenario, repensar la confianza no es solo un desafío teórico, sino una tarea urgente para el diseño de tecnologías justas, comprensibles y responsables.

3. Conclusiones: Principales hallazgos y asuntos pendientes

En los debates contemporáneos sobre la inteligencia artificial (IA) en medicina, el concepto de confianza ocupa un lugar cada vez más central. Sin embargo, a medida que intentamos pensar con mayor precisión este fenómeno, nos enfrentamos a dificultades que no se deben solo a su complejidad intrínseca, sino también a una tendencia persistente a fragmentarlo. Una de las formas más comunes de esta fragmentación consiste en intentar localizar la confianza en uno de los polos de la relación: o bien se la concibe como una actitud del sujeto —el profesional de la salud o el paciente— o bien se la atribuye a las propiedades objetivas del sistema tecnológico

ISSN-e: 1885-5679

—su precisión, su fiabilidad, su grado de explicabilidad—. Pero esta separación, aunque operativamente útil en ciertos análisis, nos parece inadecuada cuando se trata de captar la naturaleza profunda de la confianza en contextos donde los actores humanos y las tecnologías están profundamente entrelazados.

Nuestra propuesta parte de una crítica a esta disyunción. Frente a la tentación de adjudicar los diversos componentes de la confianza a uno u otro polo —lo subjetivo o lo objetivo—, sostenemos que muchas de las categorías clave que organizan este debate solo pueden ser comprendidas en un plano relacional. No se trata simplemente de reconocer que la confianza se da entre un sujeto y un sistema, sino de aceptar que ciertos constructos que parecen referirse a uno de los polos en realidad no pueden existir sino en la interacción misma. La confianza no es un atributo que se posea, como si se tratara de un capital transferible, sino un fenómeno emergente que se constituye en el acto de relación.

Esto se hace especialmente evidente cuando analizamos conceptos como responsabilidad, rendición de cuentas, antropomorfismo o alineación de valores. Todos ellos suelen ser tratados como propiedades atribuibles, pero presentan resistencias notables a esta lógica asignativa. En particular, el concepto de responsabilidad en el contexto de la IA médica muestra de forma clara los límites de una visión que separa tajantemente a humanos y máquinas como entidades independientes con funciones bien delimitadas. Decir que un sistema de IA es "responsable" en un sentido robusto resulta profundamente problemático. No solo porque carece de voluntad, de agencia moral o de capacidad para comprender las consecuencias éticas de sus actos, sino porque semejante afirmación desvincula de manera implícita a los agentes humanos de la cadena de decisión. Esto no significa que debamos volver a modelos donde el profesional de la salud asuma toda la carga de la acción clínica, ignorando las mediaciones tecnológicas, sino más bien que necesitamos un modelo de responsabilidad distribuida, en el que los humanos y los sistemas técnicos compartan, aunque de forma asimétrica, los espacios de acción, decisión y justificación.

La responsabilidad, por tanto, no puede entenderse como un atributo estático ni como un simple reparto de tareas. Se trata de una estructura dinámica que emerge en la relación práctica entre agentes humanos y sistemas automatizados. Tal como lo

ISSN-e: 1885-5679

señaló Matthias (2004), existe una zona de "lagunas de responsabilidad" (responsibility gaps) cuando los sistemas actúan de formas que no son totalmente controladas ni por los programadores ni por los usuarios. Pero incluso si aceptamos esta tesis, no deberíamos concluir que la responsabilidad ha desaparecido, sino que ha cambiado de forma: ha pasado de ser una cuestión de imputación individual a una cuestión de diseño relacional. De hecho, uno de los principales desafíos éticos de la IA médica no consiste tanto en localizar al responsable último, como en establecer estructuras relacionales que permitan una atribución de responsabilidad coherente y distribuida.

Un tercer constructo que ilustra con claridad la naturaleza relacional de la confianza es el antropomorfismo. Lejos de ser un mero fenómeno perceptivo o un sesgo cognitivo, el impulso antropomórfico cumple una función estructurante: permite a los usuarios interpretar el comportamiento del sistema en términos comprensibles, incluso cuando esas interpretaciones no se ajustan estrictamente a su funcionamiento real. Este fenómeno ha sido estudiado desde múltiples perspectivas, y si bien se lo ha criticado por inducir falsas expectativas o por fomentar una atribución indebida de capacidades morales a los sistemas, también es cierto que cumple un papel pragmático en la interacción cotidiana con tecnologías complejas. En este sentido, el antropomorfismo no es una simple proyección del usuario, sino una dimensión relacional que se co-construye en el uso, y que puede ser modulada —aunque no completamente eliminada— mediante decisiones de diseño y educación tecnológica.

En suma, defender una concepción relacional de la confianza en IA no significa disolver las distinciones entre lo humano y lo técnico, ni negar la importancia de las características individuales de cada componente. Significa, más bien, entender que muchas de las categorías que utilizamos para pensar estos problemas —responsabilidad, rendición de cuentas, antropomorfismo, alineación— no se dejan reducir a uno de los polos de la relación sin que se pierda algo esencial en el proceso. Se trata de conceptos cuya inteligibilidad depende de una estructura de interacción, de una práctica compartida, de un tejido institucional y simbólico que les da sentido. Solo al reconocer esta dimensión relacional podremos aspirar a desarrollar tecnologías verdaderamente confiables, no solo porque funcionen bien, sino porque estén insertas en relaciones humanas que puedan sostener su uso de manera crítica, responsable y justa.

ISSN-e: 1885-5679

4. Referencias bibliográficas

Choung, H., David, P., & Ross, A. (2022). Trust in AI and its role in the acceptance of AI technologies. International Journal of Human–Computer Interaction, 1–13. https://doi.org/10.1080/10447318.2022.2050543

Floridi, L. (2019). Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. 32, 185–193

Hatherley, J. J. (2020). Limits of trust in medical AI. Journal of Medical Ethics, 46(7), 478–481. https://doi.org/10.1136/medethics-2019-105935

London AJ, (2019). Artificial intelligence and black-box medical decisions: accuracy versus Explainability. Hastings Cent Rep;49(1):15–21

Malle, B. F., and Ullman, D. (2021). "A multidimensional conception and measure of human-robot trust" in Trust in human-robot interaction. (Cambridge, MA: Academic Press), 3–25.

Shin, D., and Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. Comput. Hum. Behav. 98, 277–284. doi: 10.1016/j.chb.2019.04.019

Sparrow, R., & Hatherley, J. (2019). The Promise and Perils of AI in Medicine. International Journal of Chinese & Amp; Comparative Philosophy of Medicine, 17(2), 79–109. https://doi.org/10.24112/ijccpm.171678

Thiebes, S., Lins, S., Sunyaev, A.: Trustworthy artificial intelligence. Electron. Mark. (2020). https:// doi. org/ 10. 1007/s12525- 020- 00441-4

Topol EJ. Deep medicine: how artificial intelligence can make healthcare human again. New York, NY: Basic Books, 2019.

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. CSCW 2021 - The 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing, 5(CSCW2), 1–39. https://doi.org/10.1145/3476068

ISSN-e: 1885-5679