Libro Blanco de IA en Medicina

Directores:
José Antonio Trujillo Ruiz y Cristina Gil Membrado.

Autores:
Ezequiel López Rubio, José Manuel Jerez Aragonés, Eduardo de Teresa, Manuel Jiménez Navarro, Francisco Miralles Linares, Luis E. Echarte Alonso, José Antonio Trujillo Ruiz, Cristina Gil Membrado.

Promueve:

commálaga Colegio de Médicos

commálaga health hub

Colaboran:

Fundación Unicaja

Ministerio de Ciencia, Innovación y Universidades

Cofinanciado por la Unión Europea

Agencia Estatal de Investigación

Promueve:

Colegio Oficial de Médicos de Málaga.
Commálaga Health Hub.

Colaboran:

Fundación Unicaja.

Ministerio de Ciencia, Innovación y Universidades.
Proyecto PID2022-136964NB-I00 El Derecho ante la Salud Digital, Personalizada y Robótica (SALUDPYR) financiado por MICIU/ AEI/10.13039/501100011033/ y por FEDER, UE

Diseño y maquetación: Álvaro Ruiz.

“Lo que ha unido la Medicina durante siglos que la Inteligencia Artificial no lo separe”

Sumario

1

Prefacio

1 Prefacio

Dr. José Antonio Trujillo

La inteligencia artificial (IA) ha irrumpido en la conversación pública con enorme fuerza, ocupando titulares, foros profesionales y debates institucionales. Sin embargo, esa abundancia de noticias, promesas y relatos sobre innovación no siempre ha ido acompañada de un desarrollo académico y profesional suficientemente sólido sobre su impacto real en la medicina: en la práctica clínica, en la relación médico-paciente, en la formación, en la ética, en la seguridad, en los derechos y en la responsabilidad profesional.

Este **Libro Blanco de la Inteligencia Artificial en Medicina** nace precisamente para contribuir a ese espacio que faltaba, un lugar de análisis riguroso, plural y clínicamente orientado, escrito desde la medicina y en diálogo con la tecnología, el derecho, la bioética y la universidad. Su vocación no es amplificar el ruido, sino ordenar el debate; no es rendirse al entusiasmo acrítico ni al temor estéril, sino ofrecer criterios para una integración responsable, humanista y científicamente fundada de la IA en el ámbito sanitario. El propio diseño de la obra —desde el panorama general de la IA en medicina hasta los capítulos dedicados al marco técnico-científico, la enseñanza, el profesionalismo, la relación médico-paciente y la protección de datos— refleja esa ambición de mirada transversal.

La publicación de este libro por el **Colegio de Médicos de Málaga**, a través del **Commálaga Health Hub**, en el que colaboran la **Fundación Unicaja** y el **Ministerio de Ciencia, Innovación y Universidades**. El **Commálaga Health Hub** fue concebido como una línea estratégica pionera dentro de los colegios de médicos en España, con el propósito de situar a la profesión médica en el centro de la conversación sobre innovación, salud digital e inteligencia artificial, no como espectadora, sino como protagonista responsable. Este libro es una expresión natural de esa visión, una apuesta por anticiparse, por pensar con rigor antes de improvisar, y por construir marcos útiles para los médicos, las instituciones y, sobre todo, para los pacientes.

Los autores que participan en esta obra son referentes en España en los campos que abordan. Desde la inteligencia artificial y la validación técnica hasta la docencia médica, el profesionalismo, la ética, la relación fiduciaria y el marco jurídico de datos y algoritmos, cada capítulo aporta conocimiento experto, experiencia y una perspectiva complementaria. Esa diversidad es una fortaleza, la IA en medicina no puede comprenderse desde una sola disciplina. Requiere una conversación seria entre clínicos, tecnólogos, juristas y humanistas, y este libro quiere ser precisamente ese punto de encuentro.

Estamos, probablemente, ante una de las primeras publicaciones en España que aborda la inteligencia artificial en medicina con esta amplitud, profundidad y enfoque colegial. Ojalá sea también un punto de inflexión. No porque cierre el debate, sino porque aspira a elevarlo. La IA no debe alejarnos de la medicina, sino ayudarnos a ejercerla mejor, con más capacidad analítica, pero también con más prudencia, más transparencia, más formación y más humanidad.

Ese es, en el fondo, el mensaje de este libro: la inteligencia artificial puede transformar la medicina, pero el rumbo de esa transformación seguirá dependiendo de nosotros. Si mantenemos en el centro la dignidad de la persona, el juicio clínico, la ética profesional y el compromiso con el bien del paciente, la IA no será una amenaza para la medicina, sino una oportunidad histórica para fortalecerla.

Con esa convicción ofrecemos esta obra a la profesión médica y a la sociedad. Nuestro deseo es que lo que ha unido durante siglos la Medicina, no lo separe la Inteligencia Artificial.

Dr. José Antonio Trujillo Ruiz
Autor, codirector y editor “Libro Blanco de la IA en Medicina”.
Vicepresidente Colegio de Médicos de Málaga.
Director del Commálaga Health Hub.

2

Presente y futuro de la IA en Medicina

Prof.Dr.Ezequiel López Rubio
Catedrático de Inteligencia Artificial.
Universidad de Málaga.

Resumen ejecutivo:

La inteligencia artificial (IA) se ha consolidado en la última década como una de las tecnologías con mayor impacto potencial en la medicina contemporánea. Desde los primeros sistemas expertos hasta los actuales modelos de aprendizaje profundo y las recientes arquitecturas generativas, la IA ha evolucionado paralelamente al aumento de la capacidad computacional, la disponibilidad de datos clínicos digitalizados y el desarrollo de algoritmos cada vez más sofisticados. Este capítulo ofrece un panorama general del estado actual de la IA en medicina, abordando los principales tipos de sistemas, sus aplicaciones clínicas más relevantes y las perspectivas de evolución hacia el año 2030. Se pone un énfasis especial en la IA generativa aplicada a la imagen médica y al soporte al diagnóstico, analizando tanto sus logros como sus limitaciones y desafíos. El objetivo es proporcionar una visión rigurosa y accesible que permita comprender el papel de la IA como herramienta de apoyo a la toma de decisiones clínicas, más que como sustituto del profesional sanitario. Asimismo, se plantean cuestiones abiertas para el debate colegial sobre la integración responsable de estas tecnologías en la práctica clínica y en los sistemas de salud.

Palabras clave:

Aprendizaje máquina; aprendizaje profundo; diagnóstico asistido por computador; toma de decisiones; IA generativa.

Executive Summary:

Artificial intelligence (AI) has, over the past decade, become established as one of the technologies with the greatest potential impact on contemporary medicine. From early expert systems to today's deep learning models and more recent generative architectures, AI has evolved in parallel with increases in computational capacity, the availability of digitized clinical data, and the development of increasingly sophisticated algorithms. This chapter provides an overview of the current state of AI in medicine, addressing the main types of systems, their most relevant clinical applications, and prospects for development through 2030. Particular emphasis is placed on generative AI applied to medical imaging and diagnostic support, examining both its achievements and its limitations and challenges. The aim is to offer

a rigorous yet accessible account that clarifies AI's role as a tool to support clinical decision-making rather than as a substitute for healthcare professionals. In addition, the chapter raises open questions to inform professional and institutional debate on the responsible integration of these technologies into clinical practice and health systems.

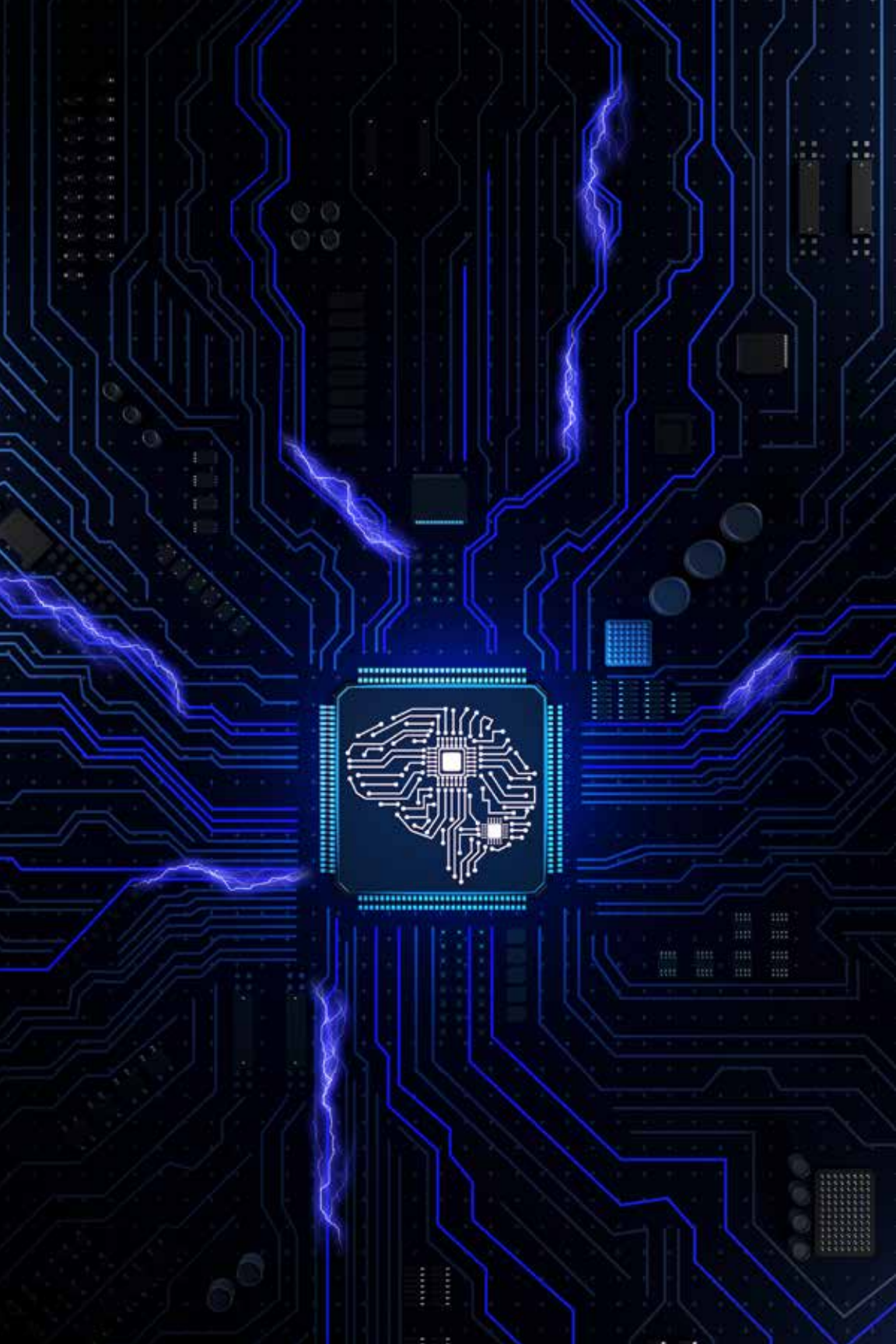
Keywords:
Machine learning; deep learning; computer-aided diagnosis; decision-making; generative AI.

Ideas fuerza:

- **La IA ya es una tecnología transversal en medicina**, impulsada por digitalización, datos clínicos y capacidad computacional.
- **Evoluciona por paradigmas** (sistemas expertos → aprendizaje automático → deep learning → IA generativa), ganando potencia pero también nuevos retos.
- **El dato es el factor crítico**: calidad, diversidad y gobernanza determinan sesgos, generalización y seguridad clínica.
- **La imagen médica es el campo más maduro**, y la IA generativa amplía capacidades (mejora/reconstrucción), con riesgos de artefactos y "alucinaciones".
- **La IA debe entenderse como apoyo a la decisión clínica**, integrada en flujos, validada y con supervisión humana y marco ético-regulatorio.

Key messages:

- **AI is now a cross-cutting technology in medicine**, driven by digitization, clinical data, and computing power.
- **It has progressed through paradigms** (expert systems → machine learning → deep learning → generative AI), increasing capability while introducing new challenges.
- **Data are the critical determinant**: quality, diversity, and governance shape bias, generalizability, and clinical safety.
- **Medical imaging is the most mature domain**, and generative AI expands capabilities (enhancement/reconstruction), with risks of artifacts and "hallucinations."
- **AI should be framed as decision support**, integrated into workflows, validated, and kept under human oversight within ethical and regulatory boundaries.



Sumario:

2 Presente y futuro de la IA en Medicina

Prof. Dr. Ezequiel López Rubio

1. Introducción: la IA en el contexto de la medicina moderna

La medicina contemporánea se encuentra inmersa en un proceso de transformación profunda impulsado por la digitalización del conocimiento clínico, la disponibilidad de datos biomédicos y el crecimiento exponencial de la capacidad de cómputo. En este escenario, la inteligencia artificial (IA) ha pasado en pocos años de ser un campo eminentemente experimental a constituir un conjunto de tecnologías con impacto real y creciente en la práctica médica cotidiana (Briganti & Le Moine, 2020; Rajpurkar et al., 2021). La IA no representa una disciplina clínica autónoma, sino un instrumento transversal que interactúa con casi todas las especialidades médicas, desde la radiología y la anatomía patológica hasta la atención primaria, la gestión sanitaria o la investigación traslacional.

De forma general, la IA puede definirse como el conjunto de métodos computacionales capaces de realizar tareas que tradicionalmente requieren inteligencia humana, tales como el reconocimiento de patrones, la predicción de eventos, el razonamiento probabilístico o el aprendizaje a partir de la experiencia. En el ámbito sanitario, estas capacidades se aplican fundamentalmente al análisis de datos clínicos complejos y heterogéneos —imágenes médicas, historias clínicas electrónicas, datos genómicos, señales fisiológicas o información procedente de dispositivos portátiles— con el objetivo de asistir al profesional en la toma de decisiones (Yin et al., 2021; Bajwa et al., 2021).

El interés creciente por la IA en medicina responde, en gran medida, a la convergencia de varios factores estructurales. En primer lugar, la sobrecarga asistencial y la creciente complejidad de la atención médica hacen cada vez más difícil que el clínico pueda integrar de forma eficiente toda la información relevante disponible. En segundo lugar, la medicina actual genera volúmenes de datos que exceden claramente la capacidad humana de análisis, pero que son especialmente adecuados para los métodos de aprendizaje automático y, en particular, de aprendizaje profundo (Rajpurkar et al., 2021; Bajwa et al., 2021). Finalmente, el envejecimiento de la población y la prevalencia creciente de enfermedades crónicas plantean retos organizativos y económicos que impulsan la búsqueda de herramientas capaces de mejorar la eficiencia, la precisión diagnóstica y la personalización de los cuidados.

Desde una perspectiva clínica, conviene subrayar que la IA no pretende sustituir al médico, sino ampliar sus capacidades. La literatura científica reciente coincide en señalar que los mayores beneficios se obtienen cuando los sistemas de IA se integran como herramientas de apoyo, actuando como "segundo lector", sistemas de alerta precoz o asistentes en la priorización de riesgos clínicos, más que como mecanismos de decisión autónoma (Rajpurkar et al., 2021; Yin et al., 2021). Esta concepción de la IA como *medicina aumentada* sitúa al profesional sanitario en el centro del proceso asistencial, manteniendo la responsabilidad clínica y el juicio médico como elementos irrenunciables.

En este marco general, la imagen médica ha emergido como uno de los ámbitos donde la IA ha alcanzado un mayor grado de madurez. La naturaleza digital de las imágenes, la disponibilidad de bases de datos y la existencia de tareas bien definidas —detección, segmentación y clasificación— han facilitado la adopción de sistemas de diagnóstico asistido por computador basados en aprendizaje profundo (Mun et al., 2021; Chan et al., 2020).

Más recientemente, la aparición de modelos de IA generativa ha ampliado este horizonte, permitiendo no solo analizar imágenes, sino también generar, mejorar o sintetizar información visual con potencial impacto en el diagnóstico, la planificación terapéutica y la formación médica (KoohiMoghadam & Bae, 2023; Rabbani et al., 2025). Estas tecnologías introducen nuevas oportunidades, pero también riesgos específicos, como la generación de artefactos (alucinaciones) clínicamente plausibles pero incorrectos, lo que refuerza la necesidad de validación rigurosa y supervisión médica.

No obstante, el despliegue de la IA en la medicina moderna plantea interrogantes relevantes que trascienden lo puramente tecnológico. Cuestiones como la interpretabilidad de los modelos, los sesgos en los datos de entrenamiento, la protección de la privacidad, la responsabilidad profesional o la adecuación del marco regulador europeo adquieren una importancia crítica en un contexto donde las decisiones asistenciales afectan directamente a la seguridad y los derechos de los pacientes (Pesapane et al., 2021; Meszaros et al., 2022). En consecuencia, cualquier aproximación rigurosa al presente y futuro de la IA en medicina debe integrar, desde el inicio, una visión clínica, ética y organizativa.

Esta introducción pretende situar al lector en ese contexto de cambio, proporcionando un marco conceptual para comprender el desarrollo posterior

del capítulo. A lo largo de las secciones siguientes se abordarán los fundamentos técnicos de la IA, sus principales tipos de sistemas y aplicaciones clínicas, con especial atención a la IA generativa en imagen médica y soporte al diagnóstico, así como los retos y oportunidades que marcarán su evolución hacia el horizonte de 2030.

2. Evolución y fundamentos de la inteligencia artificial en medicina

2.1. De los sistemas expertos al aprendizaje automático

Los primeros intentos de informatizar el razonamiento clínico se materializaron en sistemas expertos: programas que codificaban, mediante reglas explícitas, el conocimiento de especialistas para emular decisiones diagnósticas o terapéuticas. Su fortaleza residía en la transparencia —cada conclusión podía rastrearse hasta una regla "sientonces"— y en la validez dentro de dominios restringidos, cuidadosamente definidos. Sin embargo, estos sistemas sufrían de tres limitaciones estructurales: (1) la adquisición de conocimiento era costosa y lenta; (2) la mantención ante nueva evidencia clínica resultaba ardua; y (3) su generalización fuera del escenario de diseño era modesta (Briganti & Le Moine, 2020; Rajpurkar et al., 2021). La práctica clínica, dinámica y heterogénea, pronto desbordó la capacidad de reglas estáticas para capturar la variabilidad de pacientes, contextos y flujos de trabajo.

A partir de la década de 2000 y, con mayor intensidad en los 2010, la llegada de historias clínicas electrónicas y la explosión de datos multimodales favorecieron el paso hacia el aprendizaje automático (*machine learning*, ML): modelos que aprenden patrones y relaciones a partir de datos etiquetados o no etiquetados, en lugar de depender de reglas codificadas manualmente. El ML clásico —árboles de decisión, máquinas de vectores soporte, *random forests*, regresiones penalizadas— demostró ventajas claras frente a los sistemas expertos, en particular mayor robustez ante ruido y capacidad de adaptación al incorporar nueva evidencia (Yin et al., 2021; Bajwa et al., 2021). Este desplazamiento conceptual implicó un cambio de paradigma: del "conocimiento experto capturado" al "conocimiento estadístico aprendido",

con el clínico pasando de autor de reglas a curador de datos y validador de resultados.

El salto cualitativo se produciría con el aprendizaje profundo (*deep learning*, DL), especialmente en radiología y anatomía patológica. Las redes neuronales convolucionales (*convolutional neural networks*, CNN) mostraron rendimiento sobresaliente en tareas de detección, segmentación y clasificación de imágenes, superando con frecuencia a enfoques tradicionales y acercándose al nivel de expertos en problemas acotados, como la detección de lesiones o la identificación de hallazgos en cribado (Chan et al., 2020; Mun et al., 2021). La razón técnica es doble: por un lado, las CNN extraen representaciones jerárquicas directamente a partir de píxeles, evitando la ingeniería manual de características; por otro, escalan eficazmente con grandes volúmenes de datos y potencia de cómputo. Este avance se integró en la práctica a través de diagnóstico asistido por computador (*computer-aided diagnosis*, CAD) de nueva generación y sistemas de priorización de listas de trabajo que mejoran la eficiencia y la seguridad del paciente.

El tránsito de sistemas expertos a ML/DL también modificó la epistemología de la ayuda a la decisión clínica. Mientras los sistemas expertos se apoyaban en conocimiento declarativo y razonamiento simbólico, los métodos de aprendizaje se sustentan en regularidades empíricas derivadas de datos, lo que introduce desafíos de interpretabilidad y calibración. La evidencia reciente subraya que, para alcanzar impacto clínico real, los sistemas ML/DL deben integrarse en flujos asistenciales y evaluarse mediante estudios prospectivos que midan efectos sobre procesos y resultados (por ejemplo, tiempos de informe, eventos adversos, métricas de precisión y utilidad clínica), más allá del rendimiento cuantitativo en estudios retrospectivos (Yin et al., 2021; Rajpurkar et al., 2021). En otras palabras, el "éxito algorítmico" debe transformarse en beneficio clínico dentro de entornos operativos complejos.

La IA generativa ha introducido una tercera fase evolutiva que complementa el análisis predictivo: modelos capaces de crear contenido (imágenes, texto, señales) con aplicaciones emergentes en imagen médica (síntesis de datos para entrenamiento, reducción de ruido, mejora de resolución, traducción entre modalidades) y en soporte al diagnóstico (borradores de informes, resúmenes, asistentes conversacionales para educación clínica). Aunque esta sección se centra en la transición histórica, merece destacarse que la IA generativa no sustituye al ML/DL discriminativo; más bien amplía

el ecosistema de herramientas, ofreciendo datos sintéticos y mejoras de calidad que potencian el entrenamiento y la operabilidad de sistemas existentes (KoohiMoghadam & Bae, 2023; Rabbani et al., 2025). En el Cuadro 1 se puede observar una comparación entre los distintos paradigmas de la IA en Medicina.

Desde el punto de vista organizativo, el paso a ML/DL exigió nuevas competencias institucionales: gobernanza de datos, protección de la privacidad, auditoría de modelos y monitorización posterior al despliegue. En Europa, el marco regulador —encabezado por el Reglamento General de Protección de Datos (RGPD) y el Reglamento Europeo de Inteligencia Artificial (*AI Act*)— está consolidando un enfoque basado en riesgo, con requisitos de calidad de datos, transparencia y vigilancia continua para aplicaciones consideradas de alto riesgo en salud (Pesapane et al., 2021; Meszaros et al., 2022). Este entorno regulatorio refuerza la idea de que la transición tecnológica debe acompañarse de responsabilidad profesional y evaluación ética, manteniendo el juicio clínico en el centro.

En síntesis, la evolución desde los sistemas expertos hacia el aprendizaje automático y profundo ha sido impulsada por la necesidad de escalar el conocimiento clínico frente a datos crecientes y tareas complejas, y por la evidencia de que los modelos aprendidos ofrecen mejor rendimiento y adaptabilidad en dominios visuales y no visuales. No obstante, el progreso técnico debe equilibrarse con interpretabilidad, validación prospectiva, equidad y marcos de gobernanza que aseguren seguridad y utilidad clínica sostenibles (Rajpurkar et al., 2021; Yin et al., 2021).

Característica	Sistemas expertos	Aprendizaje automático clásico	Aprendizaje profundo
Base de conocimiento	Reglas explícitas definidas por expertos	Datos estructurados y variables seleccionadas	Grandes volúmenes de datos (especialmente imagen)
Tipo de razonamiento	Simbólico, lógico	Estadístico	Subsimbólico, basado en representaciones jerárquicas
Transparencia	Alta (reglas interpretables)	Media	Baja a media (modelo «caja negra»)
Necesidad de datos	Baja	Moderada	Alta
Capacidad de adaptación	Limitada	Buena	Muy alta
Ámbitos clínicos típicos	Diagnóstico en dominios cerrados	Predicción de riesgo, clasificación	Radiología, anatomía patológica, visión médica
Ejemplos	MYCIN, INTERNISTI	Modelos de riesgo clínico	CAD en mamografía, TC, RM
Principales limitaciones	Rigidez y escasa escalabilidad	Dependencia de ingeniería de variables	Interpretabilidad, sesgos, coste computacional

Cuadro 1. Comparación entre sistemas expertos, aprendizaje automático y aprendizaje profundo en medicina. Fuente: elaboración propia a partir de Briganti y Le Moine (2020), Chan et al. (2020) y Rajpurkar et al. (2021).

2.2. Aprendizaje profundo y disponibilidad de datos

Tal como se expone en el Cuadro 1, el aprendizaje profundo se distingue de los enfoques previos de inteligencia artificial por su alta dependencia de grandes volúmenes de datos y por su capacidad para aprender representaciones complejas de forma automática. Esta característica ha sido determinante para su adopción en medicina, especialmente en aquellos ámbitos donde los datos se generan de manera sistemática en formato digital, como la imagen médica y las señales biosanitarias (Chan et al., 2020; Rajpurkar et al., 2021).

El rendimiento de los modelos de aprendizaje profundo está estrechamente vinculado a la cantidad, diversidad y calidad de los datos disponibles. A diferencia del aprendizaje automático clásico, que puede operar razonablemente bien con conjuntos de datos moderados y variables cuidadosamente seleccionadas, el aprendizaje profundo necesita miles o millones de ejemplos para ajustar sus millones de parámetros internos sin incurrir en sobreajuste. En medicina, esta exigencia ha impulsado el desarrollo de grandes repositorios de imágenes radiológicas, datos histopatológicos y registros clínicos estructurados, así como iniciativas de colaboración multicéntrica para compartir datos de forma segura y regulada (Mun et al., 2021; Yin et al., 2021).

La imagen médica ha sido el terreno más fértil para el aprendizaje profundo por varias razones convergentes. En primer lugar, las imágenes constituyen un dato altamente estructurado, con patrones espaciales que pueden ser capturados eficazmente por redes neuronales convolucionales. En segundo lugar, los servicios de radiología y medicina nuclear han generado durante décadas volúmenes masivos de imágenes digitalizadas, almacenadas en sistemas PACS (*Picture Archiving and Communication System*), lo que ha permitido entrenar modelos con conjuntos de datos de tamaño suficiente para tareas como detección de lesiones, segmentación de órganos o clasificación de hallazgos (Chan et al., 2020). Esta disponibilidad explica que, tal como refleja el Cuadro 1, el aprendizaje profundo haya alcanzado mayor madurez en radiología y anatomía patológica que en otras áreas clínicas.

Sin embargo, la disponibilidad de datos no garantiza automáticamente utilidad clínica. Los datos médicos reales suelen ser incompletos, heterogéneos y sesgados, reflejando prácticas asistenciales, poblaciones específicas o limitaciones del sistema sanitario. Estos sesgos pueden trasladarse

al modelo entrenado, generando rendimientos desiguales según grupos de edad, sexo o procedencia geográfica. Diversos estudios han evidenciado que modelos entrenados en centros únicos o en poblaciones poco diversas presentan dificultades al generalizarse a otros contextos clínicos (Rajpurkar et al., 2021; Pesapane et al., 2021). Por este motivo, la calidad y representatividad del dato son elementos tan críticos como su volumen.

En este contexto, han cobrado especial relevancia estrategias para maximizar el valor de los datos disponibles, como la transferencia de aprendizaje (*transfer learning*), el aprendizaje federado y la generación de datos sintéticos. La transferencia de aprendizaje permite reutilizar modelos previamente entrenados en grandes conjuntos de datos generales, adaptándolos a tareas clínicas específicas con un volumen reducido de datos locales, lo que resulta especialmente útil en hospitales de menor tamaño. Por su parte, el aprendizaje federado posibilita el entrenamiento distribuido de modelos sin necesidad de centralizar los datos, contribuyendo a preservar la privacidad del paciente y a cumplir con los requisitos regulatorios europeos (Meszaros et al., 2022).

La emergencia de la IA generativa, abordada en capítulos posteriores, ha añadido una capa adicional a la relación entre aprendizaje profundo y disponibilidad de datos. Modelos generativos como las redes adversarias generativas (*generative adversarial networks*, GAN) y los modelos de difusión permiten crear imágenes médicas sintéticas que pueden emplearse para aumentar conjuntos de entrenamiento, equilibrar clases poco representadas o simular escenarios clínicos infrecuentes. Tal como se adelanta en el Cuadro 1, estas técnicas no sustituyen a los datos reales, pero pueden mejorar la robustez de los modelos de aprendizaje profundo cuando se utilizan con criterios metodológicos rigurosos (KoohiMoghadam & Bae, 2023; Rabbani et al., 2025).

Desde el punto de vista organizativo y regulador, la dependencia del aprendizaje profundo respecto a los datos plantea desafíos adicionales. La extracción, anonimización, almacenamiento y reutilización secundaria de datos clínicos debe realizarse conforme al Reglamento General de Protección de Datos (RGPD) y a las disposiciones del Reglamento Europeo de Inteligencia Artificial. Dichos marcos normativos exigen garantizar la trazabilidad del dato, la explicabilidad del sistema y la supervisión humana de las decisiones asistidas por IA, especialmente cuando estas se consideran de alto riesgo (Pesapane et al., 2021; Meszaros et al., 2022).

En síntesis, el aprendizaje profundo representa un punto de inflexión en la aplicación de la inteligencia artificial a la medicina, pero su éxito está indisolublemente ligado a la disponibilidad responsable de datos clínicos de alta calidad. Como sugiere el Cuadro 1, su potencia técnica debe equilibrarse con limitaciones claras en términos de interpretabilidad, coste computacional y gobernanza del dato. Comprender esta relación resulta esencial para valorar de forma realista sus aplicaciones clínicas actuales y su proyección futura en el sistema sanitario.

3. Tipos de sistemas de IA aplicados a la medicina

La aplicación de la inteligencia artificial en medicina se materializa a través de sistemas con funcionalidades diferenciadas, cuyo diseño y utilidad varían según la especialidad clínica, el tipo de dato disponible y la naturaleza de la decisión asistencial. Desde el punto de vista práctico, resulta más útil clasificar estos sistemas no solo por su arquitectura técnica, sino por el rol clínico que desempeñan en el proceso diagnóstico, terapéutico u organizativo. En las subsecciones siguientes se describen los principales tipos de sistemas de IA actualmente utilizados o en fase de implantación, contextualizados por especialidades médicas.

3.1. Sistemas predictivos y de clasificación clínica

Los sistemas predictivos utilizan modelos de aprendizaje automático para estimar la probabilidad de eventos clínicos futuros o para clasificar pacientes en grupos de riesgo. Estos sistemas se basan habitualmente en datos estructurados procedentes de historias clínicas electrónicas, analíticas, constantes vitales y antecedentes médicos.

En medicina interna y atención hospitalaria, estos modelos se emplean para la predicción de sepsis, deterioro clínico, reingresos o mortalidad intrahospitalaria. Diversos estudios han mostrado que los modelos de aprendizaje automático pueden identificar patrones sutiles de descompensación clínica con mayor antelación que las escalas tradicionales, siempre que se integren como sistemas de alerta y no como sustitutos del juicio clínico (Yin et al., 2021; Rajpurkar et al., 2021).

En cardiología, los sistemas predictivos se han aplicado a la estratificación de riesgo cardiovascular, la detección de arritmias a partir de electrocardiogramas y la predicción de insuficiencia cardíaca utilizando datos longitudinales. Los modelos basados en redes neuronales han mostrado un rendimiento comparable o superior al de puntuaciones (*scores*) clínicas clásicas en poblaciones específicas, especialmente cuando combinan señales fisiológicas y variables clínicas (Bajwa et al., 2021).

En atención primaria, estos sistemas tienen un potencial especial para apoyar la priorización de pacientes, la identificación de multimorbilidad y el cribado oportunista de enfermedades crónicas, aunque su adopción real ha sido más limitada por problemas de integración en los flujos de trabajo y por consideraciones éticas relacionadas con la equidad y el sesgo poblacional (Yin et al., 2021).

3.2. Sistemas de diagnóstico asistido por computador (CAD)

Los sistemas de diagnóstico asistido por computador constituyen uno de los ámbitos más maduros y clínicamente relevantes de la IA médica. Su función principal es apoyar al especialista en la interpretación de estudios diagnósticos, actuando como segundo lector o como herramienta de priorización.

En radiología, la IA se emplea de forma creciente en tareas como la detección de nódulos pulmonares en tomografía computarizada, la identificación de hallazgos en mamografía o la segmentación automática de lesiones en resonancia magnética. Las redes neuronales convolucionales han demostrado un alto rendimiento en tareas bien delimitadas, especialmente en contextos de cribado, donde pueden reducir la carga de trabajo y mejorar la consistencia diagnóstica (Chan et al., 2020; Mun et al., 2021).

En anatomía patológica, los sistemas de IA analizan imágenes digitalizadas de biopsias para detectar patrones histológicos asociados a malignidad, grado tumoral o marcadores pronósticos. La llamada *patología digital* ha abierto la puerta a modelos que no solo replican tareas humanas, sino que identifican características subvisuales no evidentes al microscopio tradicional (Rajpurkar et al., 2021).

En dermatología y oftalmología, especialidades con fuerte componente visual, se han desarrollado sistemas capaces de clasificar lesiones cutáneas

o retinopatía diabética con niveles de precisión elevados en estudios controlados. No obstante, la literatura subraya que su utilidad clínica real depende de su validación en entornos asistenciales variados y de su uso como apoyo, no como diagnóstico autónomo (Yin et al., 2021).

3.3. Sistemas de apoyo a la toma de decisiones clínicas

Los sistemas de apoyo a la toma de decisiones clínicas integran información procedente de múltiples fuentes para sugerir diagnósticos diferenciales, opciones terapéuticas o alertas de seguridad. A diferencia de los sistemas CAD, su alcance suele ser más transversal y orientado al proceso clínico global.

En oncología, estos sistemas se utilizan para apoyar la selección de tratamientos basados en guías clínicas, características moleculares del tumor y resultados previos en pacientes similares. Aunque algunos sistemas han sido ampliamente difundidos, la evidencia disponible indica que su valor reside principalmente en la estandarización y la actualización del conocimiento, más que en la sustitución del juicio del oncólogo (Briganti & Le Moine, 2020).

En farmacología clínica y medicina hospitalaria, la IA se emplea para detectar interacciones medicamentosas, ajustar dosis en función de variables dinámicas y prevenir eventos adversos. Estos sistemas han mostrado impacto positivo en seguridad del paciente cuando se diseñan con umbrales clínicamente razonables, evitando la fatiga por alertas (Bajwa et al., 2021).

La progresiva incorporación de modelos de lenguaje y IA generativa está ampliando estas funciones hacia la síntesis de información clínica, la redacción asistida de informes y el apoyo a la educación médica, aspectos que se desarrollarán en apartados posteriores.

3.4. Sistemas organizativos y de gestión sanitaria

Finalmente, un grupo de sistemas de IA con creciente relevancia, aunque a menudo menos visibilizado, es el destinado a la gestión de recursos sanitarios y optimización organizativa.

En servicios de urgencias, los modelos predictivos se utilizan para anticipar picos de demanda, estimar tiempos de estancia y mejorar la asignación

de camas. En gestión hospitalaria, la IA contribuye a la planificación de quirófanos, la gestión de listas de espera y el análisis de eficiencia de procesos, con el objetivo de mejorar la sostenibilidad del sistema sin comprometer la calidad asistencial (Bajwa et al., 2021).

Aunque estos sistemas no toman decisiones clínicas directas, su impacto indirecto sobre la atención al paciente es significativo, lo que justifica la necesidad de criterios de transparencia, validación y supervisión equivalentes a los aplicados en sistemas clínicos.

4. IA generativa: conceptos y relevancia clínica

4.1. ¿Qué es la IA generativa?

La inteligencia artificial generativa hace referencia a un conjunto de modelos y técnicas de aprendizaje automático diseñados no solo para analizar, clasificar o predecir a partir de datos existentes, sino para generar nuevos contenidos que reproducen las características estadísticas y estructurales de los datos de entrenamiento. A diferencia de los sistemas discriminativos tradicionales —orientados a la asignación de etiquetas o a la estimación de probabilidades—, los modelos generativos aprenden la distribución subyacente de los datos, lo que les permite crear instancias nuevas y plausibles, como imágenes, texto, señales biomédicas o datos clínicos sintéticos (Topol, 2022).

Desde el punto de vista técnico, la IA generativa se sustenta principalmente en arquitecturas avanzadas de aprendizaje profundo. Entre las más relevantes se encuentran las redes generativas adversarias (*Generative Adversarial Networks*, GAN), los autocodificadores variacionales (*Variational Autoencoders*, VAE) y, más recientemente, los modelos de difusión. Estas arquitecturas han demostrado una elevada capacidad para generar datos de alta fidelidad, especialmente en dominios complejos como la imagen médica, donde resulta imprescindible preservar tanto la coherencia anatómica como la representación realista de patrones patológicos.

En el ámbito de la medicina, la IA generativa ha cobrado especial relevancia debido a su potencial para abordar algunas de las limitaciones estructurales

de los sistemas de IA convencionales. Uno de los principales desafíos en el desarrollo de modelos clínicos robustos es la disponibilidad limitada de datos de alta calidad y correctamente anotados, en particular en enfermedades raras, poblaciones infrarrepresentadas o escenarios clínicos poco frecuentes. Los modelos generativos permiten la creación de datos sintéticos realistas que pueden emplearse para aumentar conjuntos de entrenamiento, mejorar la generalización de los algoritmos y reducir el riesgo de sobreajuste, siempre que su uso se realice bajo criterios metodológicos estrictos y con adecuada validación clínica.

Más allá de la generación de datos, la IA generativa desempeña un papel activo en aplicaciones clínicas concretas. En el ámbito de la imagen médica, estos modelos se utilizan para la mejora de la calidad de imagen, la reducción de ruido, la reconstrucción de imágenes a partir de adquisiciones incompletas o de baja dosis —como en tomografía computarizada o resonancia magnética— y la armonización de imágenes procedentes de distintos dispositivos o protocolos de adquisición. Estas aplicaciones tienen implicaciones directas en la seguridad del paciente, al permitir reducir la exposición a radiación o acortar los tiempos de exploración, así como en la eficiencia de los procesos diagnósticos.

Desde una perspectiva clínica y organizativa, resulta fundamental subrayar que la IA generativa debe entenderse como una tecnología de apoyo integrada en sistemas más amplios de diagnóstico asistido por computador y soporte a la toma de decisiones clínicas. La capacidad de generar contenidos no implica autonomía clínica ni sustitución del profesional sanitario, sino que exige supervisión, validación y contextualización por parte de expertos. Este enfoque es clave para garantizar un uso seguro, ético y clínicamente relevante de la IA generativa en los sistemas de salud contemporáneos.

4.2. Aplicaciones en imagen médica

La imagen médica constituye uno de los ámbitos donde la inteligencia artificial generativa ha alcanzado un mayor grado de desarrollo y relevancia clínica. Modalidades como la tomografía computarizada, la resonancia magnética, la ecografía o la imagen histopatológica generan grandes volúmenes de datos visuales cuya adquisición, procesamiento e interpretación plantean retos técnicos, asistenciales y organizativos. En este contexto, los modelos generativos han demostrado un notable potencial para mejorar la calidad

de imagen, optimizar los flujos de trabajo clínicos y reforzar los sistemas de diagnóstico asistido por computador (Topol, 2022).

Una de las aplicaciones más consolidadas de la IA generativa en imagen médica es la mejora y reconstrucción de imágenes. Mediante el uso de redes generativas adversarias y modelos de difusión, es posible reconstruir imágenes de alta calidad a partir de adquisiciones incompletas o de baja señal. Este enfoque resulta especialmente relevante en técnicas que implican exposición a radiación ionizante, como la tomografía computarizada, donde la IA generativa permite reducir la dosis administrada al paciente sin comprometer la calidad diagnóstica de las imágenes obtenidas.

Otra línea de aplicación de creciente interés es la armonización y normalización de imágenes procedentes de distintos dispositivos, fabricantes o protocolos de adquisición. La variabilidad técnica entre equipos de imagen constituye un obstáculo importante para la generalización de los modelos de IA y para la comparación longitudinal de estudios. Los modelos generativos permiten transformar imágenes heterogéneas en representaciones más homogéneas, facilitando tanto el entrenamiento de algoritmos robustos como su despliegue en entornos clínicos diversos.

La generación de datos sintéticos representa igualmente una contribución clave de la IA generativa al ámbito de la imagen médica. Estos datos artificiales, cuando se generan a partir de modelos validados, pueden utilizarse para ampliar conjuntos de entrenamiento, especialmente en escenarios donde el acceso a imágenes reales está limitado por restricciones éticas, legales o de privacidad. La utilización de imágenes sintéticas puede mejorar el rendimiento de los modelos diagnósticos y contribuir a reducir sesgos asociados a la infrarrepresentación de determinadas patologías o grupos poblacionales.

Desde la perspectiva clínica, la IA generativa aplicada a imagen médica se integra habitualmente como un componente de sistemas más amplios de apoyo al diagnóstico. Su función principal no es reemplazar la interpretación del especialista, sino proporcionar imágenes de mayor calidad, facilitar la identificación de regiones de interés o apoyar comparaciones longitudinales que refuercen el juicio clínico. Este enfoque es coherente con las recomendaciones actuales, que subrayan la necesidad de mantener la supervisión humana y la responsabilidad clínica en el uso de sistemas de IA en medicina.

En conjunto, la IA generativa aplicada a imagen médica representa una evolución significativa de las herramientas diagnósticas basadas en IA. Su impacto potencial abarca desde la mejora de la calidad técnica de las imágenes hasta la optimización de los procesos asistenciales y la reducción de riesgos para el paciente. No obstante, su adopción generalizada exige procesos rigurosos de validación clínica, evaluación regulatoria y formación específica de los profesionales sanitarios, factores que serán determinantes para su consolidación en la práctica clínica habitual.

4.3. Soporte al diagnóstico y formación

Además de sus aplicaciones en la mejora y generación de imágenes, la inteligencia artificial generativa desempeña un papel creciente como herramienta de soporte al diagnóstico clínico y a la formación de profesionales sanitarios. En este contexto, su valor no reside en la emisión autónoma de diagnósticos, sino en la capacidad de complementar el razonamiento clínico mediante la generación de información visual, contextual o explicativa que facilite la interpretación de hallazgos y la toma de decisiones fundamentadas (Topol, 2022).

En el ámbito del diagnóstico asistido por computador, la IA generativa puede integrarse como un componente clave en sistemas híbridos que combinan modelos discriminativos y generativos. Mientras que los primeros se orientan a la detección o clasificación de patrones patológicos, los segundos pueden contribuir a resaltar regiones de interés, generar visualizaciones alternativas o simular la evolución de determinadas lesiones bajo distintos supuestos clínicos. Este enfoque resulta especialmente útil en áreas como la radiología, la anatomía patológica digital y la dermatología, donde la interpretación visual es central para el proceso diagnóstico.

La generación de ejemplos sintéticos y escenarios clínicos simulados constituye otra aportación relevante de la IA generativa al soporte diagnóstico. Estos recursos permiten a los profesionales explorar variaciones de presentación de una misma patología, incluyendo casos poco frecuentes o atípicos, lo que puede favorecer una mayor sensibilidad diagnóstica y una mejor preparación ante situaciones clínicas complejas. Asimismo, los modelos generativos pueden emplearse para evaluar la robustez de los sistemas diagnósticos frente a variaciones en la calidad de imagen o en las características de los datos de entrada.

En el ámbito de la formación médica, la IA generativa abre nuevas posibilidades para el aprendizaje basado en simulación. La generación de imágenes clínicas realistas, casos virtuales interactivos y escenarios personalizados permite diseñar entornos formativos adaptados al nivel de experiencia del estudiante o del profesional en formación. Estas herramientas resultan especialmente valiosas en disciplinas donde el acceso a determinados casos está limitado por su baja prevalencia o por consideraciones éticas, como ocurre en algunas enfermedades raras o en procedimientos invasivos.

Desde una perspectiva pedagógica, la IA generativa puede contribuir a una formación más activa y centrada en el aprendizaje por casos, favoreciendo el desarrollo del razonamiento clínico y la capacidad de toma de decisiones. No obstante, la literatura reciente subraya la necesidad de integrar estas tecnologías dentro de programas formativos estructurados, evitando una dependencia acrítica de los sistemas automatizados y promoviendo la comprensión de sus limitaciones y posibles sesgos.

En conjunto, la IA generativa aplicada al soporte diagnóstico y a la formación sanitaria representa una herramienta prometedora para reforzar tanto la práctica clínica como la educación médica. Su impacto dependerá, en gran medida, de su integración responsable en los entornos asistenciales y educativos, de la validación rigurosa de sus resultados y de la capacitación de los profesionales para interactuar de forma crítica y competente con estos sistemas. Este enfoque resulta coherente con la concepción de la IA como un apoyo al profesional sanitario, orientado a mejorar la calidad de la atención y la seguridad del paciente, sin sustituir el juicio clínico humano.

5. Impacto clínico actual de la IA

La inteligencia artificial ha pasado, en la última década, de ser una tecnología experimental a convertirse en una herramienta con impacto tangible en múltiples ámbitos de la práctica clínica. Aunque su grado de adopción varía entre especialidades y sistemas sanitarios, existe un consenso creciente en que la IA ya está contribuyendo de forma significativa a mejorar procesos diagnósticos, optimizar flujos de trabajo y apoyar la toma de decisiones clínicas. Este impacto, no obstante, debe analizarse de manera realista, atendiendo tanto a los beneficios demostrados como a las limitaciones

observadas en su implementación en entornos asistenciales reales (Topol, 2022).

Uno de los ámbitos donde el impacto clínico de la IA es más evidente es el diagnóstico asistido por computador (*Computer-Aided Diagnosis*, CAD). Sistemas basados en aprendizaje profundo han demostrado un rendimiento comparable, y en algunos casos superior, al de expertos humanos en tareas específicas como la detección de lesiones en imágenes radiológicas, la identificación de patrones histopatológicos o el cribado de enfermedades dermatológicas.

En radiología, por ejemplo, la IA se utiliza de forma creciente para priorizar estudios urgentes, detectar hallazgos incidentales y reducir la carga de trabajo asociada a tareas repetitivas. Estos sistemas no reemplazan al radiólogo, sino que actúan como una segunda lectura automatizada que puede aumentar la sensibilidad diagnóstica y reducir la probabilidad de errores por fatiga o sobrecarga asistencial. Resultados similares se han observado en patología digital, donde la IA facilita la detección de áreas sospechosas en grandes preparaciones histológicas.

Más allá del diagnóstico, la IA está comenzando a influir en la toma de decisiones clínicas mediante sistemas de apoyo que integran información procedente de múltiples fuentes, como datos clínicos, analíticos, de imagen y genómicos. Estos sistemas pueden ayudar a estimar riesgos, predecir eventos adversos o sugerir estrategias terapéuticas basadas en patrones observados en grandes cohortes de pacientes.

En áreas como la medicina intensiva, la oncología o la cardiología, se han desarrollado modelos predictivos capaces de anticipar deterioros clínicos, complicaciones o respuestas al tratamiento. No obstante, la evidencia disponible indica que el mayor valor de estos sistemas se alcanza cuando sus recomendaciones son interpretadas dentro del contexto clínico y validadas por el profesional sanitario, evitando una automatización acrítica de decisiones complejas.

El impacto clínico de la IA no se limita a la interacción directa con el paciente, sino que también se extiende a la optimización de procesos organizativos y logísticos. Algoritmos de aprendizaje automático se utilizan para mejorar la gestión de recursos, la planificación de agendas, la asignación de camas o

la predicción de demanda asistencial, contribuyendo a una utilización más eficiente de los sistemas sanitarios.

Estas aplicaciones, aunque menos visibles desde el punto de vista clínico, pueden tener un efecto indirecto significativo sobre la calidad de la atención, al reducir tiempos de espera, mejorar la continuidad asistencial y liberar tiempo del profesional sanitario para tareas de mayor valor clínico.

A pesar de los avances descritos, el impacto clínico real de la IA sigue estando condicionado por importantes limitaciones. Muchos modelos muestran un rendimiento elevado en entornos controlados, pero experimentan una disminución de su eficacia cuando se despliegan en contextos clínicos distintos de aquellos en los que fueron entrenados. Esta falta de generalización está relacionada con la heterogeneidad de los datos, las diferencias en protocolos clínicos y la variabilidad poblacional.

Asimismo, persisten retos relacionados con la interpretabilidad de los modelos, la integración en los flujos de trabajo clínicos y la aceptación por parte de los profesionales. La evidencia sugiere que la adopción exitosa de la IA depende tanto de factores técnicos como organizativos y culturales, lo que subraya la necesidad de enfoques multidisciplinares en su implementación.

Finalmente, resulta fundamental destacar que el impacto clínico de la IA debe evaluarse mediante estudios prospectivos, ensayos clínicos y análisis de resultados en salud, y no únicamente a través de métricas técnicas. En los últimos años, se ha producido un aumento de iniciativas orientadas a establecer marcos de evaluación clínica y regulatoria más robustos, que permitan determinar de forma objetiva el valor añadido de la IA en la práctica asistencial.

En conjunto, la IA ya está generando un impacto clínico real, aunque heterogéneo, en la medicina contemporánea. Su consolidación como herramienta habitual dependerá de la capacidad para demostrar beneficios clínicos sostenibles, garantizar la seguridad del paciente y fomentar una integración responsable y centrada en el profesional sanitario.

6. Proyección de la IA en medicina hacia 2030

El horizonte de 2030 se perfila como un período de consolidación y expansión de la inteligencia artificial en medicina, con transformaciones significativas tanto en la práctica clínica como en la organización de los sistemas de salud. Se espera que la IA deje de ser una herramienta complementaria o experimental para integrarse de manera fluida en los flujos asistenciales, la educación médica y la gestión sanitaria, promoviendo una atención más personalizada, eficiente y segura.

Se anticipa que la IA se consolidará como un apoyo integral a la toma de decisiones clínicas, con sistemas capaces de procesar y sintetizar información multimodal —historial clínico, imágenes médicas, genómica, monitorización remota y datos de vida real— para ofrecer recomendaciones contextualmente relevantes. La combinación de IA generativa y modelos predictivos permitirá escenarios clínicos simulados, proyecciones de evolución de enfermedades y estimaciones de riesgo más precisas, favoreciendo decisiones individualizadas y basadas en evidencia.

Se espera un crecimiento sostenido en el uso de IA generativa, no solo en imagen médica, sino en áreas como la patología digital, la planificación quirúrgica y la telemedicina avanzada. Los modelos generativos podrían generar datos sintéticos para entrenar sistemas en patologías raras, crear simulaciones para la formación de profesionales y mejorar la precisión diagnóstica en entornos con recursos limitados. Esto permitirá reducir brechas de equidad y aumentar la capacidad de respuesta frente a escenarios clínicos complejos o emergencias sanitarias.

Hacia 2030, la IA contribuirá de forma creciente a la optimización de procesos organizativos y logísticos: gestión de recursos, predicción de demanda hospitalaria, planificación quirúrgica y coordinación de cuidados crónicos. Además, se espera un mayor enfoque en medicina preventiva y estratificación de riesgo poblacional, apoyado por análisis predictivos basados en grandes volúmenes de datos y algoritmos de aprendizaje profundo. La capacidad de anticipar eventos adversos, detectar factores de riesgo tempranos y recomendar intervenciones personalizadas puede transformar la gestión sanitaria y los programas de salud pública.

El despliegue masivo de la IA plantea desafíos significativos en cuanto a ética, privacidad, gobernanza y responsabilidad clínica. Será esencial contar con marcos regulatorios claros, estándares de validación robustos y estrategias de IA explicable que permitan a los profesionales comprender y supervisar las recomendaciones de los sistemas automatizados. Asimismo, la confianza del paciente y del profesional sanitario será un factor crítico para la adopción efectiva de estas tecnologías.

El futuro de la IA en medicina requiere un enfoque interdisciplinario, que combine informática médica, bioingeniería, ética, gestión sanitaria y formación clínica. La colaboración del humano con el sistema de IA se consolidará como el paradigma dominante, en el que la IA potencia las capacidades del profesional sanitario sin sustituir su juicio clínico. Este modelo colaborativo permitirá maximizar los beneficios clínicos y reducir riesgos asociados a decisiones automatizadas, garantizando una atención centrada en el paciente.

En resumen, hacia 2030 la IA se perfila como un componente integral de la medicina moderna, capaz de mejorar la precisión diagnóstica, optimizar la gestión sanitaria, impulsar la formación profesional y fortalecer la medicina personalizada y preventiva. Su consolidación dependerá de la validación clínica rigurosa, la regulación efectiva y la integración responsable en los sistemas de salud.

7. Conclusiones y propuestas para el debate colegial

La integración de la inteligencia artificial en la práctica clínica plantea oportunidades significativas, pero también desafíos éticos, regulatorios y organizativos que requieren la atención de colegios médicos, sociedades científicas y autoridades sanitarias. A continuación, se presentan algunas propuestas y preguntas orientadoras que pueden servir como base para discusiones colegiales, comités de ética y foros institucionales.

7.1. Propuestas

a) Desarrollo de marcos de validación clínica estandarizados:

Los colegios médicos pueden colaborar en la definición de protocolos para evaluar la seguridad, eficacia y generalización de sistemas de IA antes de su despliegue clínico, incluyendo ensayos prospectivos y estudios multicéntricos.

b) Fomento de la formación continua en IA:

Incorporar programas educativos que permitan a los profesionales sanitarios comprender los fundamentos, capacidades y limitaciones de la IA, incluyendo su aplicación en diagnóstico asistido, imagen médica y soporte a la toma de decisiones.

c) Promoción de IA explicable y auditabilidad:

Incentivar el desarrollo de sistemas cuya lógica y resultados sean interpretables, permitiendo supervisión clínica efectiva y trazabilidad en decisiones asistenciales.

d) Ética y gobernanza de datos:

Garantizar que los datos utilizados para entrenar y evaluar modelos de IA sean representativos, respeten la privacidad de los pacientes y cumplan normativas nacionales e internacionales de protección de datos.

e) Colaboración interdisciplinaria:

Establecer mesas de trabajo que integren profesionales sanitarios, ingenieros, expertos en bioética y gestores de salud para orientar la implementación de la IA en contextos clínicos y educativos.

7.2. Preguntas para el debate colegial

a) ¿Cómo garantizar que los sistemas de IA se utilicen como apoyo al profesional sanitario y no como sustituto del juicio clínico?

b) ¿Qué criterios deberían establecer los colegios médicos para certificar la seguridad y eficacia de un sistema de IA antes de su uso clínico?

c) ¿De qué manera se puede integrar la IA en la formación médica continua y en la educación universitaria sin comprometer la autonomía crítica del profesional?

d) ¿Qué estrategias deben adoptarse para minimizar sesgos de datos y garantizar equidad en los resultados de los modelos de IA, especialmente en poblaciones infrarrepresentadas?

e) ¿Cómo se puede fomentar la confianza del paciente y del profesional sanitario en los sistemas de IA, asegurando transparencia, explicabilidad y responsabilidad legal?

f) ¿Qué modelos de colaboración interdisciplinaria podrían promover una implementación ética, segura y efectiva de la IA en entornos clínicos?

Este conjunto de propuestas y preguntas busca estimular la reflexión crítica y la toma de decisiones colegial, promoviendo un enfoque de IA centrado en el profesional sanitario y en el paciente, equilibrando innovación tecnológica con seguridad clínica, ética y gobernanza institucional.

2 Presente y futuro de la IA en Medicina

Prof. Dr. Ezequiel López Rubio

Bibliografía:

1. Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: Transforming the practice of medicine. Future Healthcare Journal, 8(2), e188–e194. https://doi.org/10.7861/fhj.2021-0095

2. Briganti, G., & Le Moine, O. (2020). Artificial intelligence in medicine: Today and tomorrow. *Frontiers in Medicine, 7*, 27. https://doi.org/10.3389/fmed.2020.00027

3. Chan, H.-P., Hadjiiski, L. M., & Samala, R. K. (2020). Computer-aided diagnosis in the era of deep learning. *Medical Physics, 47*(5), e218–e227. https://doi.org/10.1002/mp.13764

4. Koohi-Moghadam, M., & Bae, K. T. Y. (2023). Generative AI in medical imaging: Applications, challenges, and ethics. *Journal of Medical Systems, 47*, 94. https://doi.org/10.1007/s10916-023-01987-4

5. Meszaros, J., Minari, J., & Huys, I. (2022). The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union. *Frontiers in Genetics, 13*, 927721. https://doi.org/10.3389/fgene.2022.927721

6. Mun, S. K., Wong, K. H., Lo, S.-C. B., Li, Y., & Bayarsaikhan, S. (2021). Artificial intelligence for the future radiology diagnostic service. *Frontiers in Molecular Biosciences, 7*, 614258. https://doi.org/10.3389/fmolb.2020.614258

7. Pesapane, F., Bracchi, D. A., Mulligan, J. F., et al. (2021). Legal and Regulatory Framework for AI Solutions in Healthcare in EU, US, China, and Russia: New Scenarios after a Pandemic. *Radiation, 1*(4), 261–276. https://doi.org/10.3390/radiation1040022

8. Rabbani, S. A., El-Tanani, M., Sharma, S., et al. (2025). Generative artificial intelligence in healthcare: Applications, implementation challenges, and future directions. *BioMedInformatics, 5*(3), 37. https://doi.org/10.3390/biomedinformatics5030037

9. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2021). AI in health and medicine. Nature Medicine, 27, 31–38. https://doi.org/10.1038/s41591-021-01614-0

10. Topol, E. J. (2022). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 28(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

11. Yin, J., Ngiam, K. Y., & Teo, H. H. (2021). Role of artificial intelligence applications in real-life clinical practice: Systematic review. *Journal of Medical Internet Research, 23*(4), e25759. https://doi.org/10.2196/25759

3

Marco técnico y científico de la IA aplicada a la medicina

Prof. Dr. José Manuel Jerez Aragonés
Catedrático de Inteligencia Artificial.
Universidad de Málaga.

Resumen

La incorporación de sistemas de IA en la práctica clínica exige algo más que buenos resultados retrospectivos: requiere evidencia de validez y utilidad en el entorno real, junto con salvaguardas que reduzcan el riesgo para pacientes, profesionales y organizaciones. Este capítulo propone un marco técnico y científico para evaluar y desplegar sistemas de inteligencia artificial (IA) en medicina con garantías de seguridad, validez y sostenibilidad. Se argumenta que el rendimiento medio en un conjunto de datos no es suficiente para justificar su uso clínico y se describe un proceso de validación por etapas que incluye evaluación retrospectiva interna, validación externa, evaluación prospectiva en condiciones reales y vigilancia post-despliegue. Se desarrollan los tres pilares de la evaluación cuantitativa —discriminación, calibración y utilidad clínica— y se introduce el concepto de "error clínicamente relevante" para alinear métricas con riesgos asistenciales. Se abordan además los retos específicos del procesamiento del lenguaje natural y de los modelos generativos, poniendo el foco en fidelidad, verificabilidad y trazabilidad para reducir errores silenciosos y confianza descalibrada. Finalmente, se integra el análisis de sesgo algorítmico y seguridad clínica como propiedades del sistema desplegado (no solo del modelo) y se conectan estos requisitos con la necesidad de un expediente técnico-clínico auditable, control de versiones y monitorización continua en el contexto regulatorio europeo.

Palabras clave

Validación, Sesgo Algorítmico, Riesgo Clínico, Explainable AI

Executive summary:

The integration of AI systems into clinical practice requires more than strong retrospective results: it requires evidence of validity and utility in real-world settings, together with safeguards that reduce risk for patients, professionals, and organizations. This chapter proposes a technical and scientific framework for evaluating and deploying artificial intelligence (AI) systems in medicine with guarantees of safety, validity, and sustainability. It argues that average performance on a dataset is not sufficient to justify

clinical use and describes a staged validation process that includes internal retrospective evaluation, external validation, prospective evaluation under real-world conditions, and post-deployment surveillance. The three pillars of quantitative evaluation—discrimination, calibration, and clinical utility—are developed, and the concept of "clinically relevant error" is introduced to align metrics with care-related risks. The chapter also addresses the specific challenges of natural language processing and generative models, focusing on fidelity, verifiability, and traceability to reduce silent errors and miscalibrated trust. Finally, it integrates the analysis of algorithmic bias and clinical safety as properties of the deployed system (not only of the model) and connects these requirements with the need for an auditable technical-clinical dossier, version control, and continuous monitoring within the European regulatory context.

Keywords:

Validation, Algorithmic Bias, Clinical Risk, Explainable AI

Ideas fuerza:

- **La IA clínica no se valida con buen rendimiento retrospectivo; se valida con evidencia de utilidad y seguridad en condiciones reales.** Un modelo puede funcionar en un dataset y fallar en la práctica clínica si cambian pacientes, procesos, documentación o flujos asistenciales.

- **La validación debe ser un proceso por etapas y continuo, no un hito puntual previo al despliegue.** La secuencia adecuada incluye evaluación retrospectiva interna, validación externa, evaluación prospectiva en entorno real y vigilancia post-despliegue.

- **No basta con medir discriminación: la calibración y la utilidad clínica son esenciales para tomar decisiones seguras.** Un sistema puede "clasificar bien" y, aun así, inducir decisiones erróneas si sus probabilidades están mal calibradas o si no aporta beneficio neto clínico.

- **En PLN y modelos generativos, la prioridad no es la fluidez del texto, sino la fidelidad, la verificabilidad y la trazabilidad.**

La salida puede ser plausible pero clínicamente incorrecta; por eso hay que medir omisiones, afirmaciones no sustentadas y errores clínicamente relevantes.

- **El sesgo algorítmico y la seguridad clínica son propiedades del sistema desplegado, no solo del modelo.**
  La seguridad real depende de barreras, factores humanos, monitorización, control de versiones y un expediente técnico-clínico auditable alineado con el marco europeo.

Key messages

- **Clinical AI is not validated by strong retrospective performance alone; it requires evidence of real-world utility and safety.**
  A model may perform well on a dataset yet fail in practice when patients, workflows, documentation, or care processes change.

- **Validation must be staged and continuous, not a one-time milestone before deployment.**
  A robust pathway includes internal retrospective evaluation, external validation, prospective real-world evaluation, and post-deployment surveillance.

- **Discrimination alone is not enough: calibration and clinical utility are essential for safe decision-making.**
  A system may classify well and still drive poor decisions if probabilities are miscalibrated or if it does not provide net clinical benefit.

- **For NLP and generative models, fluency is not the primary criterion; fidelity, verifiability, and traceability are.**
  Outputs may be plausible yet clinically wrong, so evaluation must explicitly measure omissions, unsupported claims, and clinically relevant errors.

- **Algorithmic bias and clinical safety are properties of the deployed system, not just the trained model.**
  Real-world safety depends on safeguards, human factors, monitoring, version control, and an auditable technical-clinical dossier aligned with the European regulatory framework.


AI

Sumario:

3 Marco técnico y científico de la IA aplicada a la medicina

Prof. Dr. José Manuel Jerez Aragonés

1. Introducción: Necesidad de un marco técnico-científico específico

La inteligencia artificial (IA) aplicada a la medicina ha evolucionado en pocos años desde demostraciones prometedoras en entornos controlados hacia un despliegue progresivo en escenarios clínicos reales. Este avance está impulsado por la digitalización de la asistencia, el crecimiento del volumen y variedad de datos clínicos, y la madurez de modelos de aprendizaje profundo y de modelos de lenguaje. Sin embargo, la literatura reciente es consistente en un punto: la distancia entre un modelo que "funciona" en un conjunto de datos y una herramienta útil, segura y sostenible en práctica clínica puede ser considerable. Cambian los pacientes, los procesos, los sistemas de información y los estilos de documentación, y esos cambios afectan tanto al rendimiento como al perfil de riesgo del sistema (Subasri et al., 2025). Además, la evidencia disponible sobre cómo monitorizar de forma práctica el rendimiento de sistemas de IA en el tiempo sigue siendo relativamente limitada y heterogénea, lo que refuerza la necesidad de plantear la validación como un proceso continuo y no como un hito puntual (Andersen et al., 2024).

En un contexto sanitario, donde los errores pueden traducirse en riesgo para el paciente, no basta con comunicar métricas agregadas de rendimiento. Se requiere un marco técnico y científico que evalúe de forma sistemática la validez (qué está midiendo realmente el modelo), la robustez (respuesta ante ruido, faltantes y variabilidad), la generalización/transportabilidad (desempeño fuera del entorno de desarrollo), los sesgos (diferencias entre subpoblaciones), el riesgo clínico (consecuencias asistenciales del error) y la capacidad de monitorización (degradación por deriva y cambios de práctica). En paralelo, durante los últimos años han aparecido guías metodológicas específicas para mejorar la transparencia y completitud de la evaluación de sistemas de IA en clínica, reforzando elementos esenciales para su auditabilidad y reproducibilidad. Entre ellas, TRIPOD+AI (Collins et al., 2024) ha actualizado el marco de reporte para modelos de predicción clínica que emplean aprendizaje automático, y STARD-AI (Sounderajah et al., 2025) ha extendido los criterios mínimos para estudios de exactitud diagnóstica centrados en IA. Además, DECIDE-AI (Vasey et al., 2022) ha puesto el foco en un vacío habitual: la evaluación temprana "en condiciones reales", donde emergen fallos de integración e interacción humano-IA que raramente se detectan en validaciones retrospectivas.

Este capítulo adopta, por tanto, una perspectiva eminentemente técnica y operativa: propone criterios y procedimientos para determinar si un sistema de IA está preparado para su uso clínico, qué evidencia es razonable exigir según su propósito y nivel de riesgo, y qué salvaguardas deben incorporarse para reducir daños evitables. Se desarrollan principios de validación técnica y científica (incluyendo validación externa y evaluación prospectiva cuando corresponda), estrategias para evaluar y mitigar sesgo algorítmico, un enfoque de gestión del riesgo clínico orientado a barreras de seguridad (supervisión humana, trazabilidad y mecanismos de contingencia), y un planteamiento de Explainable AI centrado en explicaciones accionables y verificables por el profesional, evitando explicabilidad "cosmética" que incremente el riesgo de confianza excesiva (Abbas et al., 2025; Freyer et al., 2024; Tun et al., 2025).

Finalmente, este marco técnico debe conectarse con la necesidad práctica de disponer de sistemas auditables y conformes en un entorno europeo cada vez más exigente. Sin entrar en cuestiones legales, la literatura reciente sobre el AI Act —el Reglamento Europeo de Inteligencia Artificial, que establece un marco común en la UE y un enfoque basado en riesgo para el desarrollo, puesta en servicio y uso de sistemas de IA— y su interacción con el ecosistema regulatorio sanitario europeo subraya que la conformidad depende, en gran medida, de que exista un expediente técnico y clínico coherente (datos, versiones, validaciones, gestión de cambios y vigilancia post-despliegue), y de que los riesgos específicos de la IA (p. ej., deriva, sesgos, falta de trazabilidad) estén controlados con medidas verificables (van Kolfschooten, 2024; Busch et al., 2024; Aboy et al., 2024).

2. Validación técnica y científica de modelos de IA en medicina

2.1. Validar en clínica y estructuración de la evidencia

Desde la perspectiva de la IA aplicada a medicina, la validación debe entenderse como un problema de "generalización bajo restricciones de seguridad". Es decir, un modelo no es "válido" porque alcance un determinado rendimiento medio, sino porque su eficacia es reproducible, transportable al entorno objetivo y operacionalmente segura cuando se integra

en un proceso asistencial. Este matiz es esencial porque, en clínica, la evaluación no puede limitarse a una caracterización estadística del error: debe incorporar el contexto de uso, el punto de decisión, el papel del profesional y el coste diferencial de los fallos. En línea con esta necesidad, en los últimos años se han desarrollado extensiones específicas de guías de reporte para estudios con IA —como CONSORT-AI (Liu et al., 2020) y SPIRIT-AI (Rivera et al., 2020)— que, más allá del formalismo editorial, reflejan un consenso metodológico: sin una descripción completa de entradas/salidas, interacción humano-IA, manejo de errores y condiciones de integración, la evidencia es difícilmente interpretable y, en la práctica, poco auditable.

Una forma técnica y operativa de ordenar la evidencia, y evitar sesgos optimistas derivados de evaluar siempre "en condiciones de entrenamiento del modelo", es plantear un proceso de validación por etapas que diferencie con claridad (i) evaluación retrospectiva interna, (ii) validación retrospectiva externa y (iii) evaluación prospectiva previa a la adopción a escala. La evaluación interna retrospectiva permite depurar problemas clásicos del aprendizaje supervisado en entornos clínicos: fugas de información, inconsistencias de etiquetado, contaminación temporal o artefactos del preprocesamiento. La validación externa retrospectiva, por su parte, estima la transportabilidad, es decir, qué ocurre cuando se despliega el modelo en otro centro o población, o, por ejemplo, cambia el período de tiempo de seguimiento de los pacientes o el estilo de la fuente documental. En términos de aprendizaje estadístico, esta etapa aproxima el comportamiento del modelo ante cambios en la distribución conjunta de entrada-salida, que en sanidad es más la norma que la excepción (Riley et al., 2024).

Cuando el sistema se integra en procesos asistenciales, la evaluación retrospectiva deja de ser suficiente, ya que pueden aparecer fallos que no se detectan en offline evaluation (latencias, errores de integración, cambios de la entrada por rediseños de la historia clínica, ambigüedades documentales y, de forma crítica, efectos de interacción humano-IA). Precisamente para cubrir este vacío se propuso DECIDE-AI (Vasey et al., 2022) como guía para la evaluación clínica temprana "en condiciones reales" de sistemas de apoyo a decisiones basados en IA, enfatizando factores humanos y riesgos de implementación que suelen quedar fuera del retrospectivo.

La etapa final, a menudo infravalorada, es la vigilancia post-despliegue. Un sistema clínico no opera en un dominio estacionario: cambian guías, terapias, perfiles de pacientes, prácticas de codificación y plantillas de documenta-

ción. Por ello, desde un punto de vista técnico, el despliegue debe asumirse como el inicio de un régimen de monitorización y control del sistema, con indicadores para detectar degradación y criterios para recalibrar, reentrenar o retirar. Revisiones recientes han sistematizado métodos propuestos y utilizados para monitorizar el rendimiento de sistemas de IA clínica, y discuten los argumentos (a favor y en contra) de distintas aproximaciones de seguimiento continuo (Andersen et al., 2024). En esa misma línea, se han descrito enfoques para la detección de cambios y degradación del rendimiento tras el despliegue, con discusión explícita sobre señales, métricas y procedimientos de vigilancia (Schinkel et al., 2023).

Este flujo de actuación por etapas también ayuda a abordar un problema ampliamente descrito en evaluaciones metodológicas de modelos clínicos: una proporción importante de estudios presenta riesgo de sesgo por decisiones de diseño (muestras insuficientes, mala gestión de faltantes, validación interna débil o ausencia de validación externa) (Navarro et al., 2021). En consecuencia, resulta claramente razonable recomendar el uso de guías de reporte como listas de verificación internas. Para modelos de predicción clínica, TRIPOD+AI (Collins et al., 2024) actualiza el marco de reporte e integra explícitamente métodos de aprendizaje automático; para estudios de exactitud diagnóstica centrados en IA, STARD-AI (Sounderajah et al., 2025) establece criterios mínimos para informar con completitud y transparencia. En el ámbito de imagen médica, CLAIM (Tejani et al., 2024) proporciona una estructura práctica para describir población objetivo, datos, entrenamiento, validación y reproducibilidad, precisamente para facilitar interpretación y comparabilidad.

2.2. Medición de lo importante: discriminación, calibración y utilidad clínica

La validación técnica de un modelo clínico debe distinguir con claridad tres propiedades: discriminación, calibración y utilidad. Este enfoque no es una preferencia terminológica; es una necesidad para evitar errores de interpretación frecuentes cuando se presentan métricas “globales” como si capturasen el valor clínico. Un modelo puede discriminar adecuadamente —es decir, ordenar pacientes por riesgo— y, sin embargo, resultar inseguro si sus probabilidades no son fiables o si conduce a decisiones ineficientes cuando se aplican umbrales reales en el flujo asistencial.

La discriminación es la propiedad más reportada y, por ello, la más propensa a sobre-interpretación. Métricas como AUC-ROC, AUC-PR, sensibilidad, especificidad o F1 son informativas, pero deben leerse en relación con prevalencia, desbalance y punto de decisión. En problemas con eventos raros, la AUC-ROC puede permanecer "aceptable" mientras el valor predictivo positivo sea insuficiente para guiar acciones sin sobrecarga, por ejemplo, por exceso de falsos positivos. De forma análoga, en extracción de información desde texto clínico, un F1 promedio puede ocultar fallos de alta gravedad clínica —negación, temporalidad, dosis— que tienen poco peso estadístico, pero un coste asistencial desproporcionado.

En muchos casos de uso clínico, la propiedad decisiva es la calibración, porque conecta la salida con la acción. Un sistema de predicción que genera probabilidades mal calibradas induce decisiones sistemáticamente sesgadas: sobre-intervención si sobreestima el riesgo, o infra-intervención si lo infraestima. Existen evaluaciones sistemáticas recientes de métodos de calibración aplicados a modelos probabilísticos en aprendizaje automático, con revisión de literatura y comparación de enfoques en condiciones controladas (Ojeda et al., 2023). Más aún, en entornos clínicos reales la calibración tiende a deteriorarse con el tiempo por cambios en poblaciones y prácticas, lo que refuerza que su evaluación no debe limitarse al desarrollo, sino integrarse en la vigilancia post-despliegue (Schinkel et al., 2023). La implicación práctica es clara: para modelos que se usan con umbrales —alertas, priorización, activación de protocolos— la calibración debe formar parte del núcleo de la validación y del mantenimiento del sistema.

La tercera capa es la utilidad clínica, entendida como el balance neto entre beneficios y daños cuando el modelo se incorpora a una decisión. Aquí es donde muchas validaciones fallan por quedarse en una lectura "clasificatoria". La utilidad no se demuestra solo "acertando más", sino mostrando que el modelo mejora decisiones a lo largo de un rango razonable de umbrales, considerando las consecuencias clínicas. La decision curve analysis (DCA) y el concepto de beneficio neto se han consolidado como herramientas interpretables para conectar modelos predictivos con decisiones clínicas y preferencias (Piovani et al., 2023). En definitiva, cuando el modelo influye en la toma de decisiones, la validación debe incorporar alguna evaluación orientada a consecuencias, no solo a capacidad discriminativa.

Para trasladar esta idea a la práctica, resulta útil introducir el concepto de "error clínicamente relevante". Desde una perspectiva técnica, el error

agregado no describe bien el perfil de riesgo porque no todas las equivocaciones tienen el mismo coste. Por eso, además de métricas globales, conviene definir —con clínicos— un conjunto reducido de fallos inaceptables (omisiones críticas, negaciones mal interpretadas, incongruencias temporales, atribuciones incorrectas) y medir su frecuencia y circunstancias. Esta aproximación hace que la validación sea más interpretable y, sobre todo, más conectada con seguridad del paciente y con decisiones reales.

Finalmente, la evaluación por subgrupos no debe tratarse como un añadido; es parte de la validación técnica en dominios heterogéneos. Si el rendimiento difiere por edad, sexo, comorbilidad, idioma/variantes lingüísticas, centro o calidad documental, el sistema distribuye su riesgo de manera desigual. Herramientas como PROBAST (Wolff et al., 2019), concebidas para evaluar riesgo de sesgo y aplicabilidad en estudios de modelos predictivos, pueden utilizarse como soporte metodológico para auditorías internas, ayudando a identificar debilidades frecuentes antes del despliegue.

2.3. Validación específica para PLN y modelos generativos

Los sistemas basados en procesamiento del lenguaje natural (PLN) y, en particular, los modelos generativos y LLMs, introducen un reto técnico adicional: su salida puede ser lingüísticamente plausible y, sin embargo, no estar plenamente sustentada por la evidencia disponible o contener imprecisiones clínicamente relevantes. Esto obliga a reorientar la validación, pasando de evaluar principalmente la "calidad superficial" del texto a evaluar de forma prioritaria la fidelidad y la verificabilidad. Desde un punto de vista de ingeniería de evaluación, el primer paso es separar con claridad el tipo de tarea, porque cada familia requiere evidencia y métricas distintas: extracción estructurada, generación asistida y recuperación/síntesis (por ejemplo, con RAG).

En extracción estructurada (entidades, eventos, relaciones y temporalidad), las métricas clásicas (precisión, exhaustividad, F1) siguen siendo necesarias, pero rara vez suficientes. En clínica, el error que importa se concentra en fenómenos lingüísticos específicos: negación ("no presenta…"), temporalidad ("hace tres meses…", "en 2021…") y atribución ("según informe externo…", "refiere el paciente…"). Un sistema puede "acertar mucho" y fallar precisamente en esos casos, con alto impacto clínico. Por ello, la validación debería reportar explícitamente rendimiento y fallos en esos ejes, además de métricas promedio.


MPR
WL: 1048 WW: 4096 [D]
MPR
WL: 1048 WW: 4096 [D]
MPR
WL: 1048 WW: 4096 [D]

En generación asistida (por ej. resúmenes de notas clínicas), el criterio fundamental es la fidelidad a la fuente: que lo generado esté sustentado por el registro y que no introduzca afirmaciones no soportadas o inferencias presentadas como hechos. La literatura reciente sobre evaluación de LLMs en salud describe, por un lado, una gran heterogeneidad de métricas y, por otro, la necesidad de flujos robustos de evaluación humana, precisamente porque las métricas automáticas de similitud no capturan adecuadamente ni la seguridad clínica ni la utilidad (Tam et al., 2024). En este sentido, la validación debería medir, de forma explícita, la tasa de afirmaciones no sustentadas, la tasa de omisiones clínicas relevantes y la consistencia temporal.

Este foco conecta con un riesgo ampliamente discutido en la literatura: la alucinación en LLMs y sus implicaciones en medicina. Un trabajo reciente orientado a práctica clínica revisa causas y estrategias de mitigación, y propone un marco para evaluar e integrar LLMs críticamente en entornos clínicos, enfatizando que la gestión del riesgo no puede descansar en impresiones subjetivas de "calidad del texto" (Roustan et al., 2025). Desde una perspectiva técnica, una salvaguarda clave es exigir que el sistema facilite verificación: idealmente, que cada afirmación clínica relevante pueda rastrearse a evidencia explícita del registro (o, si es inferencia, que se marque como tal). Este requisito, además, mejora la auditabilidad y reduce el riesgo de confianza descalibrada.

Cuando se emplean arquitecturas de recuperación y generación (RAG), la validación debe cubrir dos componentes: (i) calidad de recuperación (qué se trae del repositorio y con qué cobertura/precisión) y (ii) fidelidad de la síntesis (si el modelo resume sin distorsionar). En estos sistemas, la trazabilidad a fragmentos fuente no es un detalle de interfaz, sino una medida de seguridad. Además, cuando el sistema se integra en flujo real —por ejemplo, en documentación asistida o apoyo a la revisión—, la evaluación debe incorporar factores humanos: carga de verificación, confianza, cambios en hábitos de documentación y efectos sobre el tiempo de trabajo clínico. En este punto, DECIDE-AI (Vasey et al., 2022) resulta especialmente pertinente al estructurar cómo reportar y evaluar, en etapas tempranas, la interacción humano-IA y los fallos operativos que emergen en condiciones reales.

Un último aspecto, a menudo olvidado, es la reproducibilidad. En PLN/LLMs, el sistema puede cambiar aunque "el modelo" no cambie: se modifican prompts, plantillas, reglas de post-proceso o el corpus de recuperación. Si no se controlan versiones, la evaluación pierde significado. En consecuencia,

una validación técnicamente sólida debe registrar versiones de modelo, prompts, fuentes de recuperación y conjuntos de evaluación, y establecer criterios de aceptación y revalidación tras cambios.

En conjunto, una validación técnica y científica "suficiente" para un sistema de IA en medicina debe (i) definir con precisión el caso de uso, el punto de decisión y el coste diferencial del error; (ii) apoyarse en un proceso por etapas que separe evaluación interna, validación externa y evaluación prospectiva cuando exista integración en flujo; (iii) reportar discriminación, calibración y, cuando el sistema conduce la toma decisiones, alguna evaluación de utilidad orientada a consecuencias; (iv) analizar rendimiento y calibración por subgrupos relevantes, incorporando marcos de evaluación del riesgo de sesgo y aplicabilidad; (v) definir y medir errores clínicamente relevantes; y (vi) para PLN/LLMs, medir explícitamente fidelidad, omisiones y afirmaciones no sustentadas, apoyándose en trazabilidad y verificación (Tam et al., 2024; Roustan et al., 2025). Finalmente, es importante concluir que la validación no concluye en el despliegue: requiere monitorización continua y gestión del ciclo de vida para detectar deriva y degradación de rendimiento.

3. Riesgos, sesgo algorítmico y seguridad clínica

3.1. Del rendimiento a la seguridad: una taxonomía de riesgos clínicos

Desde una perspectiva técnica, un sistema de IA en medicina debe tratarse como un componente socio-técnico: su comportamiento observable no depende solo del modelo, sino del entorno de datos, del flujo asistencial, de la interfaz y de la interacción con profesionales. Por ello, el riesgo clínico no se reduce al error estadístico medio, sino a cómo se distribuyen y materializan los fallos en pacientes, tiempos y contextos. Este planteamiento conecta con marcos que enfatizan la necesidad de evaluación temprana en condiciones reales y de incorporar factores humanos y de implementación en la evaluación de sistemas de soporte a decisiones basados en IA (Vasey et al., 2022).

Una taxonomía útil, suficientemente técnica y operativa, distingue los riesgos a tres diferentes niveles. En el nivel del paciente, el riesgo principal es el

daño derivado de falsos negativos (no detectar/alertar) o falsos positivos (intervenciones innecesarias, sobretratamiento, ansiedad, sobrecarga del sistema), así como de recomendaciones o resúmenes incorrectos que condicionen decisiones. En el nivel del profesional, aparecen riesgos de automatización y delegación acrítica, degradación de habilidades, fatiga por alertas o aumento de carga de verificación si el diseño no está bien resuelto; en este punto, la evidencia disponible sobre confianza y uso de sistemas de soporte basados en IA refuerza que la adopción segura requiere confianza calibrada, no confianza ciega (Tun et al., 2025). En el nivel organizativo, destacan la falta de trazabilidad y auditoría, los fallos de integración y continuidad de servicio, y la dependencia de cadenas de suministro (modelos, actualizaciones, cambios de versión) que pueden modificar el comportamiento del sistema sin que el usuario lo perciba.

Este enfoque tripartito ayuda a formalizar una idea clave: en clínica, la seguridad rara vez se garantiza "solo con más AUC". Se garantiza con un conjunto de barreras y con un proceso de gestión del ciclo de vida que asume explícitamente el cambio de condiciones (drift), la heterogeneidad del dato y la inevitabilidad de fallos. En particular, la monitorización post-despliegue se está consolidando como un requisito práctico para detectar degradación, cambios de calibración y señales de deriva, con recomendaciones recientes sobre qué medir y cómo operacionalizar esa vigilancia en sistemas clínicos (Andersen et al., 2024; Schinkel et al., 2023).

Un modo técnico de aterrizar el análisis es adoptar un razonamiento similar al de la ingeniería de seguridad: identificar modos de fallo, su severidad, su detectabilidad y la capacidad de mitigación. Para sistemas de predicción/alerta, los modos de fallo típicos incluyen degradación por cambio de población, sesgos de medición (variables registradas de forma distinta en distintos servicios) y cambios de práctica. Para sistemas generativos o de documentación asistida, emergen modos de fallo adicionales: errores silenciosos (texto plausible pero incorrecto), atribución errónea, omisiones relevantes y alucinaciones; de ahí la importancia de exigir verificabilidad y trazabilidad de las afirmaciones relevantes en el contexto clínico (Tam et al., 2024; Roustan et al., 2025). Aunque el detalle metodológico puede desarrollarse en apartados posteriores, aquí interesa fijara claramente la consecuencia: el "riesgo clínico" debe considerarse una propiedad del sistema desplegado, no solo del modelo entrenado.

3.2. Sesgo algorítmico: fuentes, medición y mitigación

El sesgo algorítmico en salud suele presentarse como una cuestión ética; sin embargo, desde un punto de vista técnico y clínico es, ante todo, un problema de validez externa y seguridad. Si un sistema aprende patrones que reflejan inequidades en acceso, calidad de documentación o decisiones previas, puede amplificarlas; y si su eficacia cae en determinados grupos, el riesgo se concentra precisamente donde la vulnerabilidad puede ser mayor. Una perspectiva influyente en la literatura médica reciente subraya que el sesgo puede originarse en múltiples puntos del flujo clínico —adquisición del dato, variabilidad demográfica, variabilidad de etiquetado, entre otros— y que, por ello, la mitigación requiere intervenciones precisas (Chen et al., 2023).

Conviene distinguir cuatro fuentes de sesgo que aparecen de manera recurrente en aplicaciones clínicas: (1) sesgo de representación (quién está en el dato), cuando las cohortes no reflejan a la población donde se desplegará el sistema; (2) sesgo de medición (cómo se registra), especialmente relevante cuando variables clínicas dependen del proceso asistencial y, en texto libre, del estilo y completitud de la documentación; (3) sesgo de etiqueta (qué se considera "verdad"), donde la variabilidad interprofesional o entre centros se traslada al modelo; y (4) sesgo de despliegue (cómo cambia el sistema), porque el sistema puede alterar comportamiento clínico y documental, retroalimentando los datos futuros. En términos técnicos, estas fuentes se traducen en cambios en la distribución de entrada o en la relación entrada-salida, y en cambios inducidos por la propia intervención algorítmica.

En los últimos años se ha consolidado un cuerpo de revisiones que ordenan métricas y técnicas de mitigación de sesgo y justicia en sanidad, con especial foco en datos de vida real. En particular, revisiones recientes sintetizan métricas de equidad y estrategias de mitigación cuando se utiliza real-world data en dominios sanitarios, y ponen de relieve que "ser justo" no es un único criterio, sino un conjunto de compromisos entre definiciones de equidad, objetivos clínicos y restricciones operativas (Huang et al., 2024).

Desde una perspectiva técnica, hay dos riesgos frecuentes al abordar equidad: (i) medir una única métrica de equidad como si fuese definitiva y (ii) medirla solo en fase de desarrollo. El primero es problemático porque diferentes definiciones (p. ej., paridad de tasas de falsos negativos, calibración por grupo) no siempre pueden satisfacerse simultáneamente; el

segundo porque el desempeño y la equidad pueden degradarse tras el despliegue. Este último punto ha motivado trabajos recientes sobre fairness drift (deriva de equidad), que enfatizan la necesidad de incorporar evaluaciones sistemáticas por subpoblaciones en la monitorización post-despliegue, ya que la equidad puede deteriorarse aunque las métricas globales aparenten estabilidad (Davis et al., 2025).

La implicación práctica es que la evaluación de sesgo debe ser doble. Primero, medición pre-despliegue: rendimiento y calibración por subgrupos relevantes, con justificación clínica de por qué esos subgrupos importan y cómo se han definido. Segundo, vigilancia post-despliegue: auditorías periódicas de desempeño y calibración por subgrupos, y mecanismos para investigar desviaciones. Esta visión no es solo teórica: en escenarios de despliegue real, el sesgo puede materializarse no únicamente en el rendimiento del modelo, sino también en procesos asistenciales y resultados si el sistema modifica priorizaciones, tiempos de respuesta o patrones de uso.

En definitiva, la mitigación puede requerir (a) intervención en datos (mejor cobertura de grupos, etiquetado más consistente, mejora de documentación), (b) intervención en el entrenamiento (ponderaciones, restricciones), (c) intervención en umbrales y reglas de decisión (para equilibrar costes clínicos por subgrupo), y (d) intervención en el despliegue (supervisión humana, diseños de interfaz que reduzcan confianza descalibrada y circuitos de revisión). En conjunto, la equidad debe tratarse como una propiedad que se diseña, se mide y se mantiene durante el ciclo de vida del sistema, no como una verificación puntual.

3.3. Seguridad clínica como propiedad del sistema: barreras, factores humanos y MLOps

La seguridad en IA clínica se consigue combinando tres elementos: barreras de seguridad (diseño del sistema), factores humanos (interacción y adopción) y operación controlada (monitorización y gestión de cambios). Esta tríada es especialmente importante porque muchos fallos relevantes en clínica son sistémicos: no ocurren porque el modelo “sea malo”, sino porque se despliega sin controles suficientes, con integración deficiente o sin procesos de reevaluación.

Barreras de seguridad y diseño “verificable”. En la práctica clínica, el diseño debe facilitar el ejercicio de la responsabilidad profesional, no sustituirlo. En

sistemas predictivos, esto suele implicar priorización y alertas con umbrales definidos, explicaciones orientadas a la acción y mecanismos para gestionar la fatiga por alertas. En sistemas generativos y de documentación asistida, la barrera más importante es la trazabilidad: que el profesional pueda verificar rápidamente en qué evidencia se basa la salida, y detectar omisiones o errores silenciosos. Este enfoque conecta con la necesidad de diseñar sistemas que favorezcan una confianza calibrada y eviten confianza descalibrada inducida por salidas plausibles pero incorrectas (Rosenbacke et al., 2024; Tun et al., 2025). Además, la lógica de auditoría rutinaria (calibración, deriva, rendimiento por subgrupos, calidad de alertas) se alinea con la recomendación de tratar el despliegue como el inicio de un ciclo de vigilancia, no como el final de la evaluación (Andersen et al., 2024).

Factores humanos: colaboración humano-IA. Un riesgo técnico-clínico clásico es asumir que la salida del modelo se usa "tal cual" en la decisión. En realidad, los clínicos interactúan con el sistema: aceptan, ignoran, reinterpretan o cambian su conducta documental. Por eso, la seguridad se ve afectada por confianza, carga cognitiva, claridad de la interfaz y por cómo se presenta la incertidumbre. La evidencia sobre confianza y adopción en sistemas de soporte clínico basados en IA refuerza que el rendimiento del modelo es condición necesaria, pero no suficiente, para efectividad y seguridad: la interacción y el diseño de la experiencia de uso condicionan tanto la utilidad como el perfil de riesgo (Tun et al., 2025).

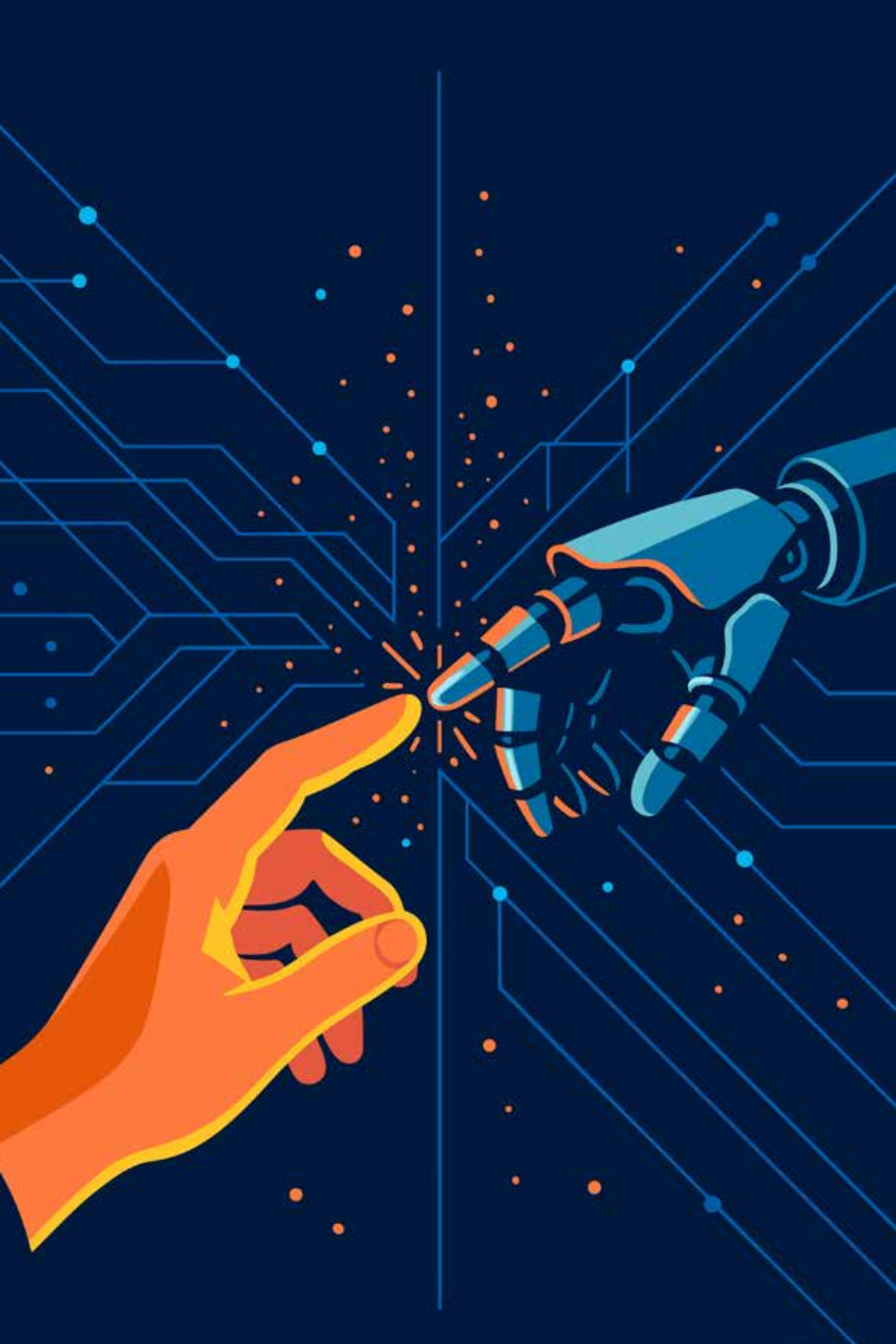
Operación y control del ciclo de vida (MLOps clínico). Desde el punto de vista de ingeniería, un modelo clínico desplegado es un sistema que requiere control de versiones, trazabilidad de datos y parámetros, monitorización de rendimiento y calibración, y gestión de cambios. La literatura sobre monitorización post-despliegue ha crecido notablemente, proponiendo enfoques para detectar degradación, sesgos emergentes y cambios de calibración, y discutiendo qué métricas son más sensibles a deriva en condiciones reales (Andersen et al., 2024; Schinkel et al., 2023). En el ámbito europeo, las revisiones recientes han subrayado la importancia de clarificar y operacionalizar la vigilancia post-mercado en dispositivos médicos con componentes de IA, y cómo las obligaciones regulatorias se conectan con prácticas de monitorización, documentación y gestión de cambios (Cuocolo et al., 2025).

En términos prácticos, hay tres decisiones operativas que determinan la seguridad en el tiempo. Primero, qué se monitoriza y con qué periodicidad: además de métricas globales, deben monitorizarse calibración, distribución

de variables clave, calidad de entrada, tasas de alerta y rendimiento por subgrupos. En particular, la evidencia reciente sobre fairness drift refuerza que la equidad puede requerir monitorización específica, no solo auditoría puntual (Davis et al., 2025). Segundo, qué desencadena una intervención: umbrales de degradación, señales de drift, cambios de prevalencia o cambios de práctica (p. ej., nueva guía, nuevo sistema de información) deben estar vinculados a acciones predefinidas —recalibración, reentrenamiento, revisión humana ampliada o retirada temporal— con trazabilidad y registro. Tercero, cómo se gobierna el cambio: debe existir un responsable clínico-técnico y un circuito de aprobación, con control de versiones y evidencia de revalidación. El enfoque de "comité de seguridad" con registros de auditoría rutinaria es una forma natural de institucionalizar esta gobernanza y sostenerla en el tiempo (Andersen et al., 2024).

Desde la teoría del aprendizaje, el problema no es que el modelo "cambie por sí solo", sino que el mundo cambia y el modelo deja de aproximar bien la relación entrada-salida. Por tanto, la seguridad clínica exige tratar la generalización como una propiedad dinámica, y la operación del sistema como una disciplina continua, no como un hito previo al despliegue (Schinkel et al., 2023; Andersen et al., 2024).

Finalmente, la evaluación del sesgo algorítmico y la seguridad clínica no son capítulos separados: son dos caras de la misma exigencia de validez en el mundo real. La literatura converge en tres mensajes prácticos: (i) las métricas globales son insuficientes si no se analizan subgrupos y calibración; (ii) la equidad puede degradarse tras el despliegue (fairness drift), por lo que requiere vigilancia; y (iii) la seguridad depende tanto del diseño de barreras y trazabilidad como de factores humanos y de procesos de monitorización y gobernanza (Davis et al., 2025; Tun et al., 2025; Andersen et al., 2024).



4. Explainable AI (XAI) y trazabilidad

4.1. La explicabilidad como requisito técnico y clínico

En medicina, la explicabilidad no debe entenderse como un adorno comunicativo para "hacer el modelo más aceptable", sino como un requisito técnico que conecta directamente con seguridad clínica, rendición de cuentas y control del riesgo. Un sistema puede mostrar buen rendimiento promedio y, aun así, fallar de manera inaceptable en subpoblaciones, en escenarios de datos incompletos o por cambios de práctica. En ese contexto, la explicabilidad aporta dos capacidades relevantes desde una perspectiva de ingeniería clínica: (i) facilitar verificación y supervisión humana (¿tiene sentido el resultado en este caso?), y (ii) apoyar auditoría y mantenimiento del sistema (¿por qué cambió el comportamiento?, ¿hay deriva de datos o de calibración?). Revisiones recientes en el ámbito sanitario han sistematizado tanto los métodos de XAI como sus limitaciones, subrayando que "explicar" en salud requiere evaluar no solo la forma de la explicación, sino su efectividad para apoyar tareas reales y evitar confianza descalibrada (Jung et al., 2023; Abbas et al., 2025).

Este punto es especialmente importante en sistemas de soporte a la decisión clínica. La literatura más reciente muestra que la adopción no depende únicamente del rendimiento, sino de una confianza calibrada: profesionales que confían lo suficiente para usar el sistema, pero no tanto como para delegar sin verificación. En este equilibrio influyen la carga de trabajo, la integración en flujo, la alineación con el juicio clínico y, de forma consistente, la disponibilidad de explicaciones comprensibles y accionables (Rosenbacke et al., 2024; Tun et al., 2025). En otras palabras, la explicabilidad no busca "convencer", sino habilitar una colaboración humano-IA segura.

Ahora bien, conviene ser técnicamente precisos: en salud existe una expectativa extendida de que la XAI revelará "cómo funciona por dentro" el modelo para múltiples perfiles (clínicos, pacientes, gestores, evaluadores). Sin embargo, trabajos recientes señalan que esta expectativa suele ser excesiva: muchas explicaciones son aproximaciones locales o visualizaciones que no equivalen a causalidad ni garantizan robustez. Por eso, un enfoque responsable consiste en formular la explicabilidad como un conjunto de funciones con objetivos concretos: apoyar verificación a nivel de caso,

facilitar auditoría de comportamiento global, y documentar límites de uso (Jung et al., 2023; Abbas et al., 2025).

4.2. Tipos de explicabilidad en la práctica clínica y evaluación

Desde una perspectiva técnica, es útil diferenciar entre tres niveles de explicación que a menudo se confunden:

Explicabilidad intrínseca: modelos cuyo funcionamiento es interpretable por diseño (p. ej., modelos lineales o reglas). Su ventaja es la transparencia; su limitación es que pueden no capturar relaciones complejas.

Explicabilidad post-hoc local: explica por qué el modelo "predijo X "en un caso concreto (p. ej., atribuciones de variables).

Explicabilidad post-hoc global: describe el comportamiento medio o patrones del modelo (p. ej., importancia de variables global, reglas aproximadas, auditorías de sensibilidad).

Las revisiones más recientes sobre XAI en sistemas de soporte a la toma de decisión muestran que, en la práctica, predominan métodos post-hoc; en imagen médica son frecuentes mapas de activación y mecanismos de atención, mientras que en datos tabulares y clínicos estructurados dominan técnicas de atribución como SHAP o LIME (Abbas et al., 2025). Aunque este patrón es comprensible porque permite añadir explicaciones sin cambiar el modelo, introduce riesgos metodológicos bien conocidos: explicaciones inestables (cambian ante pequeñas perturbaciones), explicaciones plausibles pero no necesariamente fieles al razonamiento del modelo, o visualizaciones que inducen una sensación de "comprensión" que no se traduce en decisiones más seguras (Jung et al., 2023; Abbas et al., 2025).

Por ello, la pregunta relevante no debería ser "¿qué método de XAI usamos?", sino "qué propiedad clínica debe cumplir la explicación". Una revisión reciente enfocada en "propiedades esenciales" y en la efectividad de explicaciones en salud (Jung et al., 2023) propone evaluar si la explicación es comprensible para el usuario objetivo, si es consistente, si mejora la detección de fallos y si evita confianza descalibrada. En paralelo, la evidencia sobre adopción y confianza en sistemas de ayuda a la toma de decisión con IA refuerza que una explicación puede incrementar la confianza sin necesa-

riamente mejorar el uso seguro, y que el objetivo deseable es una confianza calibrada: suficiente para apoyar el uso, pero no para delegar sin verificación (Rosenbacke et al., 2024; Tun et al., 2025). Esto refuerza una recomendación técnica clara: la explicabilidad debe validarse como parte de la intervención, no asumirse como un beneficio automático (Jung et al., 2023).

En términos operativos, una explicación clínicamente útil suele cumplir tres criterios: (a) accionabilidad: debe ayudar a responder "¿qué debo hacer ahora?" o "¿qué debo comprobar?", no solo "qué variables influyeron". Por ejemplo, en un sistema de alerta, una explicación útil puede destacar señales principales y, a la vez, indicar condiciones que vuelven el resultado menos fiable (datos faltantes, valores atípicos, contexto no representado); (b) verificabilidad: la explicación debe permitir contrastar rápidamente el resultado con la evidencia disponible. Esto es especialmente importante cuando el sistema trabaja con texto clínico o genera contenido (donde la explicación debe estar anclada a fragmentos concretos de las fuentes de documentos), de manera que la revisión humana sea eficiente y segura; y (c) honestidad epistemológica: debe dejar claro qué es evidencia y qué es inferencia, y evitar presentar como causal lo que es correlacional. En clínica, donde la atribución causal es especialmente delicada, esta distinción es esencial para no inducir interpretaciones erróneas.

La trazabilidad es, en este marco, el puente entre XAI y seguridad. En sistemas de ayuda a la toma de decisión basados en IA, la trazabilidad incluye registro de versión del modelo, datos de entrada relevantes, salida generada y —cuando aplica— referencias a la información utilizada (p. ej., qué notas o informes sustentan una afirmación). Este registro soporta auditoría y gestión de incidencias y, además, ayuda a construir confianza calibrada porque permite al profesional verificar en vez de "creer". La literatura reciente insiste en que, sin integración adecuada en interfaz y flujo, incluso buenas técnicas de explicación pierden efectividad y no resuelven por sí solas los problemas de adopción y seguridad (Abbas et al., 2025; Tun et al., 2025).

5. Certificación y conformidad en el marco europeo

5.1. Certificación de un algoritmo clínico

En el marco europeo, hablar de certificación de algoritmos clínicos suele ser una simplificación de un proceso más amplio: demostrar conformidad con requisitos aplicables, apoyándose en un sistema de gestión de calidad, gestión de riesgos, evidencia clínica y vigilancia post-comercialización. En la práctica, cuando la IA se incorpora a un producto que entra en el ámbito de dispositivo médico, el punto de partida es el Reglamento (UE) 2017/745 (MDR), que establece obligaciones sobre requisitos generales de seguridad y funcionamiento, sistema de gestión de calidad, evaluación clínica y post-market surveillance (PMS). En paralelo, la literatura reciente ha insistido en que, en productos digitales regulados, el foco no recae solo en "el modelo", sino en el sistema completo y su ciclo de vida (Aboy et al., 2024; Busch et al., 2024).

La llegada del Reglamento (UE) 2024/1689 (AI Act) añade una capa horizontal basada en riesgo para sistemas de IA y, en particular, para los considerados de "alto riesgo". Este reglamento introduce requisitos adicionales orientados a riesgos específicos de la IA (por ejemplo, gobernanza de datos, documentación técnica, transparencia, supervisión humana, robustez y monitorización), complementando marcos sectoriales como MDR/IVDR cuando el sistema de IA se integra en un dispositivo médico. La literatura especializada subraya precisamente esa lógica de aplicación simultánea y complementaria: MDR/IVDR cubren riesgos y obligaciones propios del dispositivo y su evaluación clínica, mientras que el AI Act refuerza obligaciones específicas de sistemas de IA y su gobernanza, especialmente relevantes en el ámbito sanitario (van Kolfschooten, 2024; Busch et al., 2024; Aboy et al., 2024).

Desde un punto de vista técnico, esto implica que el "objetivo" del expediente no es solo justificar una métrica de rendimiento, sino demostrar, de forma auditable, tres propiedades del sistema desplegado:

Seguridad y funcionamiento previstos: el sistema hace lo que declara, dentro de límites de uso y población objetivo, con un perfil de riesgo aceptable.

Control del ciclo de vida: se conoce qué versión del sistema está en uso, cómo se gestiona el cambio (incluido reentrenamiento o modificaciones de prompts/datos) y cómo se revalida.

Vigilancia y mejora continua: existe un sistema PMS capaz de detectar degradación, incidentes y señales de sesgo o deriva, y de accionar medidas correctoras.

Estas tres propiedades conectan directamente con el enfoque de ingeniería que requiere la IA clínica: en entornos no estacionarios, la conformidad no puede descansar en una validación puntual, sino en trazabilidad, control de versiones y monitorización. En particular, la literatura reciente sobre vigilancia post-mercado en dispositivos médicos con IA ha enfatizado la necesidad de operacionalizar PMS con mecanismos de retroalimentación clínica y procedimientos de actualización controlada, precisamente para sostener seguridad y rendimiento en condiciones reales (Cuocolo et al., 2025).

5.2. El expediente auditable

Sin entrar en cuestiones normativas, resulta útil traducir la exigencia de conformidad a un conjunto mínimo de artefactos técnicos que permitan evaluar, auditar y mantener un sistema de IA en clínica. En términos operativos, el expediente puede organizarse en cinco bloques de documentos y evidencias auditables que hacen explícitos el uso previsto, el perfil de riesgo, la eficacia del sistema, el control del cambio y la vigilancia post-despliegue, de modo que la evaluación no dependa solo de métricas aisladas, sino de un sistema trazable y gobernable.

(1) Definición de uso previsto y límites de uso

El expediente debe dejar inequívoco el propósito clínico (diagnóstico, priorización, apoyo documental, etc.), el contexto (servicio, punto del flujo), el tipo de entrada y salida, y los límites (poblaciones excluidas, datos mínimos, condiciones donde el sistema no debe utilizarse). Este bloque es crítico porque condiciona todo lo demás: métricas relevantes, umbrales operativos, evidencia clínica exigible y perfil de riesgo. En particular, la literatura reciente sobre el encaje regulatorio europeo en productos digitales enfatiza que la definición de uso previsto y límites de uso no es un formalismo: es el ancla

técnica de la evaluación y de la vigilancia posterior (Aboy et al., 2024; Busch et al., 2024).

(2) Gestión del riesgo y requisitos de seguridad

En MDR/IVDR, la gestión del riesgo es un eje central y se integra con requisitos de seguridad y funcionamiento; además, la vigilancia post-comercialización forma parte del sistema de calidad y del mantenimiento del producto. En IA, este bloque debe incluir: modos de fallo esperables (p. ej., degradación por cambios de distribución, errores por entradas incompletas, sesgos), severidad y detectabilidad, y barreras (supervisión humana, verificación, fallback). El valor añadido del AI Act —cuando aplica, especialmente en sistemas de alto riesgo— es exigir controles explícitos sobre riesgos característicos de IA (robustez, gobernanza de datos, supervisión humana, documentación técnica), complementando los marcos sectoriales sanitarios (van Kolfschooten, 2024; Busch et al., 2024; Aboy et al., 2024).

(3) Evidencia de eficacia: validación técnica y evaluación clínica

En marcos europeos, esta evidencia se conecta con la evaluación clínica del dispositivo y con la demostración de seguridad y funcionamiento en el uso previsto. En términos metodológicos, guías recientes como TRIPOD+AI y STARD-AI ofrecen un marco útil para estructurar y documentar esta evidencia de forma consistente y auditable (Collins et al., 2024; Sounderajah et al., 2025). En escenarios donde el sistema entra en flujo real, marcos como DECIDE-AI ayudan a estructurar la evaluación temprana prospectiva y a capturar riesgos de implementación e interacción humano-IA que el retrospectivo no revela (Vasey et al., 2022).

(4) Gestión del cambio: versiones, actualizaciones y revalidación

En IA, el cambio puede ser estructural (nuevo modelo) o sutil (cambios en datos, prompts, corpus de recuperación en RAG, umbrales, o incluso plantillas de HCE). Para un expediente auditable, el criterio técnico es simple: si cambia algo que puede cambiar el comportamiento, debe quedar registrado y debe activar un procedimiento de reevaluación proporcional al riesgo. Esta disciplina se alinea con el enfoque de “sistema” y con la necesidad de trazabilidad, robustez y supervisión que subraya el AI Act cuando corresponde (van Kolfschooten, 2024; Aboy et al., 2024). En la

práctica, también se vincula a estrategias de monitorización y detección de deriva post-despliegue (Andersen et al., 2024; Schinkel et al., 2023).

(5) Vigilancia post-comercialización y monitorización post-despliegue (PMS/PMM)

Este es el bloque que más diferencia un prototipo de un sistema clínico "sostenible". La vigilancia post-mercado no debe verse como un requisito administrativo, sino como una parte técnica esencial del control del riesgo: en entornos no estacionarios, la seguridad depende de detectar degradación, deriva, cambios de calibración y sesgos emergentes. Revisiones recientes han sistematizado aproximaciones de monitorización en IA clínica y han discutido su implementación práctica (Andersen et al., 2024; Schinkel et al., 2023). Además, en el contexto europeo, se ha enfatizado la necesidad de operacionalizar vigilancia post-mercado específicamente para dispositivos médicos con IA, conectando obligaciones regulatorias con prácticas concretas de monitorización, documentación y gestión de cambios (Cuocolo et al., 2025).

Para IA, un PMS técnicamente sólido debería incluir, como mínimo: i) monitorización de distribución de entradas (señales de drift); ii) monitorización de métricas clínicas relevantes (incluida calibración); iii) auditorías por subgrupos cuando el riesgo de inequidad sea material (incluida vigilancia de fairness drift cuando proceda) (Davis et al., 2025); iv) seguimiento de tasas de alerta/uso y señales de fatiga o confianza descalibrada; y v) registro de incidentes y circuito de corrección (incluida retirada/rollback).

Este bloque conecta también con el AI Act cuando aplica, ya que incorpora obligaciones de seguimiento, documentación y control de riesgos para sistemas de alto riesgo (van Kolfschooten, 2024; Aboy et al., 2024).

Bibliografía

1. Abbas Q, Jeong W, Lee SW. Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges. Healthcare (Basel). 2025 Aug 29;13(17):2154. doi: 10.3390/healthcare13172154. PMID: 40941506; PMCID: PMC12427955.

2. Aboy M, Minssen T, Vayena E. Navigating the EU AI Act: implications for regulated digital medical products. NPJ Digit Med. 2024 Sep 6;7(1):237. doi: 10.1038/s41746-024-01232-3. PMID: 39242831; PMCID: PMC11379845.

3. Andersen ES, Birk-Korch JB, Hansen RS, Fly LH, Röttger R, Arcani DMC, Brasen CL, Brandslund I, Madsen JS. Monitoring performance of clinical artificial intelligence in health care: a scoping review. JBI Evid Synth. 2024 Dec 1;22(12):2423-2446. doi: 10.11124/JBIES-24-00042. PMID: 39658865; PMCID: PMC11630661.

4. Busch F, Kather JN, Johner C, Moser M, Truhn D, Adams LC, Bressem KK. Navigating the European Union Artificial Intelligence Act for Healthcare. NPJ Digit Med. 2024 Aug 12;7(1):210. doi: 10.1038/s41746-024-01213-6. PMID: 39134637; PMCID: PMC11319791.

5. Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, Sahai S, Mahmood F. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng. 2023 Jun;7(6):719-742. doi: 10.1038/s41551-023-01056-8. Epub 2023 Jun 28. PMID: 37380750; PMCID: PMC10632090.

6. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;385:e078378. doi: 10.1136/bmj-2023-078378. Erratum in: BMJ. 2024 Apr 18;385:q902. doi: 10.1136/bmj.q902. PMID: 38626948; PMCID: PMC11019967.

7. Cuocolo R, Bernardini D, Pinto Dos Santos D, Klontzas ME, Akinci D'Antonoli T, Semedo LC, Decoster R, Huisman M, Kotter E, Martí-Bon-

matí L, Minoiu C, Neri E, Nikolaou K, Radzina M, Sala E, Shelmerdine SC, Topff L, Williams MC; European Society of Radiology (ESR). AI medical device post-market surveillance regulations: consensus recommendations by the European Society of Radiology. Insights Imaging. 2025 Dec 12;16(1):275. doi: 10.1186/s13244-025-02146-8. PMID: 41385025; PMCID: PMC12701188.

8. Davis SE, Dorn C, Park DJ, Matheny ME. Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. J Am Med Inform Assoc. 2025 May 1;32(5):845-854. doi: 10.1093/jamia/ocaf039. PMID: 40079820; PMCID: PMC12012346.

9. Freyer N, Groß D, Lipprandt M. The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons. BMC Med Ethics. 2024 Oct 1;25(1):104. doi: 10.1186/s12910-024-01103-2. PMID: 39354512; PMCID: PMC11443763.

10. Huang Y, Guo J, Chen WH, Lin HY, Tang H, Wang F, Xu H, Bian J. A scoping review of fair machine learning techniques when using real-world data. J Biomed Inform. 2024 Mar;151:104622. doi: 10.1016/j.jbi.2024.104622. Epub 2024 Mar 6. PMID: 38452862; PMCID: PMC11146346.

11. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon. 2023 May 8;9(5):e16110. doi: 10.1016/j.heliyon.2023.e16110. PMID: 37234618; PMCID: PMC10205582.

12. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ. 2020 Sep 9;370:m3164. doi: 10.1136/bmj.m3164. PMID: 32909959; PMCID: PMC7490784.

13. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KGM, Hooft L. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ. 2021 Oct 20;375:n2281. doi: 10.1136/bmj.n2281. PMID: 34670780; PMCID: PMC8527348.

14. Ojeda FM, Jansen ML, Thiéry A, Blankenberg S, Weimar C, Schmid M, Ziegler A. Calibrating machine learning approaches for probability estimation: A comprehensive comparison. Stat Med. 2023 Dec

20;42(29):5451-5478. doi: 10.1002/sim.9921. Epub 2023 Oct 17. PMID: 37849356.

15. Piovani D, Sokou R, Tsantes AG, Vitello AS, Bonovas S. Optimizing Clinical Decision Making with Decision Curve Analysis: Insights for Clinical Investigators. Healthcare (Basel). 2023 Aug 10;11(16):2244. doi: 10.3390/healthcare11162244. PMID: 37628442; PMCID: PMC10454914.

16. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020 Sep;26(9):1351-1363. doi: 10.1038/s41591-020-1037-7. Epub 2020 Sep 9. PMID: 32908284; PMCID: PMC7598944.

17. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, Bonnett LJ, Collins GS. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024 Jan 15;384:e074820. doi: 10.1136/bmj-2023-074820. PMID: 38224968; PMCID: PMC10788734.

18. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. JMIR AI. 2024 Oct 30;3:e53207. doi: 10.2196/53207. PMID: 39476365; PMCID: PMC11561425.

19. Roustan D, Bastardot F. The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. Interact J Med Res. 2025 Jan 28;14:e59823. doi: 10.2196/59823. PMID: 39874574; PMCID: PMC11815294.

20. Schinkel M, Boerman AW, Paranjape K, Wiersinga WJ, Nanayakkara PWB. Detecting changes in the performance of a clinical machine learning tool over time. EBioMedicine. 2023 Nov;97:104823. doi: 10.1016/j.ebiom.2023.104823. Epub 2023 Oct 2. PMID: 37793210; PMCID: PMC10550508.

21. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, Golub RM, Denniston AK, Shetty S, Moher D, Bossuyt PM, Darzi A, Ashrafian H; STARD-AI Steering Committee. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat

Med. 2025 Oct;31(10):3283-3289. doi: 10.1038/s41591-025-03953-8. Epub 2025 Sep 15. PMID: 40954311.

22. Subasri V, Krishnan A, Kore A, Dhalla A, Pandya D, Wang B, Malkin D, Razak F, Verma AA, Goldenberg A, Dolatabadi E. Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models. JAMA Netw Open. 2025 Jun 2;8(6):e2513685. doi: 10.1001/jamanetworkopen.2025.13685. PMID: 40465297; PMCID: PMC12138723.

23. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, Osterhoudt H, Wu X, Visweswaran S, Fu S, Mathur P, Cacciamani GE, Sun C, Peng Y, Wang Y. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. 2024 Sep 28;7(1):258. doi: 10.1038/s41746-024-01258-7. PMID: 39333376; PMCID: PMC11437138.

24. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, Kahn CE Jr; CLAIM 2024 Update Panel. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. Radiol Artif Intell. 2024 Jul;6(4):e240300. doi: 10.1148/ryai.240300. PMID: 38809149; PMCID: PMC11304031.

25. Tun HM, Rahman HA, Naing L, Malik OA. Trust in Artificial Intelligence-Based Clinical Decision Support Systems Among Health Care Workers: Systematic Review. J Med Internet Res. 2025 Jul 29;27:e69678. doi: 10.2196/69678. PMID: 40772775; PMCID: PMC12440830.

26. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, Denniston AK, Faes L, Geerts B, Ibrahim M, Liu X, Mateen BA, Mathur P, McCradden MD, Morgan L, Ordish J, Rogers C, Saria S, Ting DSW, Watkinson P, Weber W, Wheatstone P, McCulloch P; DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022 May 18;377:e070904. doi: 10.1136/bmj-2022-070904. PMID: 35584845; PMCID: PMC9116198.

27. van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): Implications for healthcare. Health Policy. 2024 Nov;149:105152. doi: 10.1016/j.healthpol.2024.105152. Epub 2024 Sep 7. PMID: 39244818

4

La Enseñanza de la Inteligencia Artificial en la Profesión Médica

Dr. Eduardo de Teresa
Catedrático Emérito de Cardiología y Director del Capítulo Hipocrático de la Sociedad Española de Cardiología.

Dr. Manuel Jiménez Navarro
Catedrático de Cardiología y Director del Departamento de Medicina de la Universidad de Málaga.

Resumen ejectutivo:

La Inteligencia Artificial (IA) se ha incorporado cada vez a más aspectos de la práctica médica. Desde el punto de vista de la enseñanza, ofrece una serie de ventajas tanto para docentes como para alumnos. Pero es preciso que los que van a utilizarla sean capaces de comprender sus fortalezas y las oportunidades que ofrece, así como los riesgos y amenazas que supone. La formación en IA debe contemplarse como esencial en el curriculum de las facultades de Medicina, y este formación debe acompañarse del entrenamiento en técnicas que permitan ahondar en el aspecto humano de la profesión. Para ello deberá adaptarse el contenido curricular de los programas de grado, reduciendo los contenidos que en el futuro inmediato no sean necesarios. Es preciso también implementar estrategias para formar tutores en este campo, así como establecer metodologías de formación en la aplicación de la IA a la clínica, tanto para residentes como para médicos en ejercicio. En este último apartado los Colegios de Médicos pueden jugar un importante papel.

Palabras clave

Inteligencia artificial médica, Educación médica, Humanismo médico, Razonamiento clínico.

Executive summary:

Artificial Intelligence (AI) is increasingly being incorporated into more aspects of medical practice. From an educational perspective, it offers a range of advantages for both instructors and students. However, it is essential that those who will use it are able to understand its strengths and the opportunities it provides, as well as the risks and threats it entails. AI training should be regarded as an essential component of medical school curricula, and it should be accompanied by training in techniques that deepen the human dimension of the profession. To this end, the curricular content of undergraduate programs should be adapted, reducing content that will not be necessary in the immediate future. It is also necessary to implement strategies to train tutors in this field, as well as to establish training methodologies for the clinical application of AI, both for residents and for practicing physicians. In this latter area, Medical Colleges/Medical Associations can play an important role.

4 La enseñanza de la inteligencia artificial en la profesión médica

Keywords

Medical artificial intelligence, Medical education, Medical humanism, Clinical reasoning.

Ideas fuerza:

- La IA debe ser formación esencial y transversal, integrada por fases a lo largo del grado.
- La IA educativa útil se parece más a un tutor socrático que a un buscador: guía el razonamiento y refuerza pensamiento crítico.
- Riesgos clave: alucinaciones y pérdida/no adquisición de habilidades; hay que prevenir *deskilling, never skilling y automation* bias con supervisión y verificación.
- El humanismo médico debe reforzarse como contrapeso estratégico: empatía, comunicación y juicio ético siguen siendo nucleares e insustituibles.
- Hay que evaluar y acreditar el "cómo" se usa la IA (procesos y competencias), no solo el resultado: ECOEs con IA, análisis crítico y evaluación en entorno real.

Key messages:

- AI must be an essential, longitudinal competency integrated in phases throughout medical training.
- Effective educational AI should act more like a Socratic tutor than a search engine, guiding reasoning and strengthening critical thinking.

- Core risks include hallucinations and skill erosion/non-acquisition; training must prevent deskilling, never-skilling, and automation bias through supervision and verification.

- Medical humanism must be strengthened as a strategic counterweight: empathy, communication, and ethical judgment remain central and irreplaceable.

- Assessment and accreditation should focus on how AI is used (process and competencies), not only outputs: AI-integrated OSCEs, critical case analysis, and workplace-based evaluation.

4 La enseñanza de la inteligencia artificial en la profesión médica

Sumario:

1. Introducción

Las aplicaciones basadas en inteligencia artificial (IA) en distintos campos, incluyendo la Medicina, llevan años entre nosotros. Sin embargo, el gran público fue consciente de la revolución que la IA suponía a raíz del lanzamiento del primer chatbot, GPT 3.5, en Noviembre de 2022. Las implicaciones en la enseñanza fueron percibidas de inmediato, y despertaron reacciones no siempre meditadas. Apenas unos meses después de su lanzamiento, a comienzos de 2023, el Distrito Escolar de Los Ángeles, en Estados Unidos, prohibió la utilización de GPT por los estudiantes. Las autoridades escolares de otros estados siguieron su ejemplo, así como países como Francia, Australia, la India...Se llegó a comparar el riesgo que los LLM *(Large Language Models)* suponían para la enseñanza con los de la reciente pandemia de CoVid. Hubo quien dijo que "Para su sorpresa y consternación, [los profesores] descubrirán que su aula ha dado positivo en GPT." La razón de este temor era que los estudiantes podrían emplear GPT, y los LLM que le siguieron, para redactar sus trabajos, lo que supondría una merma en su capacidad para alcanzar ciertas habilidades.

En el lado contrario, el de las empresas que impulsaban esas herramientas, también se había percibido el interés por su aplicación a la enseñanza, aunque esta vez intentando darle un enfoque positivo. En el verano de 2022, antes del lanzamiento público del primer GPT, el presidente y el CEO de OpenAI, Greg Brockman y Sam Altman, se pusieron en contacto con Salman Khan, fundador de la Khan Academy, una conocida institución sin animo de lucro dedicada a la enseñanza. Esta academia facilita cursos online gratuitos sobre las más diversas materias, desde niveles de enseñanza primaria hasta universitarios, a más de 150 millones de usuarios en todo el mundo. La finalidad era explorar formas de colaboración entre ambas empresas.

También el posible impacto de la IA en la enseñanza de la Medicina fue intuido de forma precoz. Según decía al poco de aparecer GPT Bernard Chang, Decano de Educación Médica de la Facultad de Medicina de la Universidad de Harvard, "Quizá cada pocas décadas tiene lugar una verdadera revolución en la forma en que enseñamos a los estudiantes de medicina y en lo que esperamos que puedan hacer cuando se conviertan en médicos. Esta es una de esas veces." (Tabla I)

Tabla I

Fortalezas de la IA	Limitaciones de la IA
Educación Médica	
La IA generativa puede facilitar la educación personalizada Puede enseñar habilidades técnicas	No puede replicar la empatía o el modelo de atención cercana a la cabecera del enfermo Es esencial la supervisión de la enseñanza
Interpretación de pruebas diagnósticas	
La adquisición e interpretación automática reducen la carga de trabajo	Los médicos son responsable legales de los informes generados por IA Es necesaria la supervisión en el diseño y entrenamiento del modelo
Toma de decisiones clínicas	
Los modelos de IA resumen una gran cantidad de información de la literatura médica, que ayuda y guía al médico	La lógica de las decisiones de la IA es opaca Los médicos deben evaluar los posibles sesgos o falsificaciones

No cabe duda, pues, que el impacto potencial que la IA suponía sobre la enseñanza fue detectado desde sus primeros pasos, y ha despertado miedos y esperanzas a todos los niveles.

Pese a ello, la formación específica en IA para médicos y estudiantes de Medicina no suele ser uno de los temas que suscitan más interés en los simposios generales sobre IA. Como ejemplo, la reunión SAIL 2025 *(Symposium on Artificial Intelligence for Learning Health Systems)*, pese al título, limita sus referencias docentes a cómo aprenden los sistemas de salud, no sus protagonistas.

2. IA aplicada a la enseñanza

La IA puede ayudar en la enseñanza a todos los niveles. La citada Khan Academy desarrolló, conjuntamente con OpenAI, un modelo específico diseñado para la docencia, Khanmigo. Esta herramienta puede actuar como un tutor personalizado para cada alumno, ayudándole a resolver dudas, detectando puntos de mejora, e incluso utilizando un método que los autores denominan socrático, en base a preguntas y respuestas, que induce al estudiante a desarrollar un pensamiento crítico. Khanmigo funciona además como ayuda para los docentes, diseñando programas y lecciones, al mismo tiempo que les permiten acceder a las interacciones del modelo con los alumnos. Los principios en que se basa este modelo, que por ahora no está diseñado para la enseñanza médica, son:

- Tutorización socrática (no da respuestas finales)
- Andamiaje progresivo
- Control del error (no “alucina” soluciones)
- Alineación curricular

Estos principios, en el caso de la enseñanza de la medicina, supondrían priorizar el entrenamiento del razonamiento clínico.


MONITOR 001
97.58
44.82
MONITOR 002
SCAN
CT
CO
R.A.
L.A.
R.V.
L.V.
STATS ANALYSIS & DATABASE
35.56
91.9
29.28
10.3
31.29
64.72
CONTROL CENTER
FILES & DATA
DENSITY SCANS
EXIT

Aunque este es el modelo más avanzado diseñado con fines docentes, existen otros que pueden aplicarse a la enseñanza de la Medicina. Los LLM generales podrían ajustarse como tutores socráticos, con el objetivo no de acertar un diagnóstico, sino de aprender a razonar. Entre los distintos modelos, podría emplearse GPT con *prompting* pedagógico, o Claude, muy bueno para razonamiento estructurado. De hecho, hay universidades que están adaptando LLM internos con *prompts* pedagógicos, bases clínicas controladas y sin acceso libre a internet. Tal es el caso de MedClinR1, desarrollado por la Universidad de Cornell, en Nueva York, o ClinTeach, de la Universidad de Hong Kong

Los LLM pueden también emplearse como pacientes virtuales y adaptarse a los ECOEs (Evaluación de Competencias Objetivas Estructuradas). También pueden emplearse en todo tipo de simulaciones. MedSimAI es un simulador de pacientes desarrollado con las universidades de Yale, UCSF, Cornell y Ohio State. La idea clave es que la IA educativa en Medicina no debe parecerse a Google, sino a un buen adjunto clínico que hace preguntas incómodas y no deja que el alumno tome una mala decisión sin darse cuenta.

Las múltiples posibilidades que la aplicación de la IA a la docencia médica supone, requieren una clara política institucional por parte de las distintas universidades, para homogeneizar los métodos y minimizar los riesgos de un diseño mal dirigido o sesgado, pero también, y esto es muy importante, la formación del profesorado. Uno de los problemas en este campo es que, por razones simplemente biográficas, gran parte de la plantilla docente está menos familiarizada con la IA que sus alumnos.

3. Formación en IA en las facultades de medicina (pregrado)

La enseñanza de Medicina en el Pregrado pretende preparar a los futuros médicos para la práctica de la medicina. Durante siglos se conocía perfectamente lo que esto exigía, pues no era de esperar que el acervo de conocimientos de que se disponía para tal fin variara de forma sustancial en el tiempo que duraban los años de estudio. La situación ha cambiado, y de forma llamativa en los últimos años. El ritmo al que se producen avances en los distintos campos técnicos de la Medicina o en otros cercanos, que condicionarán la forma en que se ejercerá la medicina dentro de muy pocos años -como es el caso de la IA- obliga a un ejercicio constante de previsión para ajustar la docencia a escenarios que aún no conocemos, que cambian de forma acelerada y que hacen muy difícil establecer previsiones manejables.

Hay, no obstante, algunas líneas que parecen claras, o al menos en las que coinciden diversos autores. Glatter, Papadakos y Shah así lo reconocen: "La formación en humanidades, ciencias sociales y comunicación es tan importante para los futuros médicos como la anatomía y fisiología. Esto es también cierto para los residentes y otros profesionales sanitarios. Por supuesto que igualmente importante es la formación sobre las nuevas tecnologías, como la IA". Parece, por tanto, que cara al futuro:

1. Es necesario incorporar una formación reglada en IA y en su aplicación a la práctica médica.
2. Es necesario incluir una formación en humanismo médico.

3.1. Formación en IA

Hemos comentado la necesidad de formar al profesorado en IA, para lo que sería necesario establecer los cursos o programas adecuados que, dado el impacto que la IA puede tener en la práctica clínica y los posibles riesgos que ello comporta, deberían ser obligatorios. Asselbergs ha resumido los componentes de la formación en IA que serían aconsejables a los distintos niveles (Tabla II)

Tabla II

Componentes esenciales de la formación en IA para médicos y estudiantes		
Área de formación	**Contenido/conocimientos**	**Dirigido a**
Fundamentos de la IA	Bases de la IA, algoritmos de machine learning y deep learning	Estudiantes
AI en diagnóstico y toma de decisiones	Cómo puede ayudar la IA en el diagnóstico, toma de decisiones y planificación terapéutica	Estudiantes, residentes
Reconocimiento y evitación de sesgos	Identificación y corrección de sesgos, asegurando la exactitud y justicia	Estudiantes, formadores
Colaboración en la toma de decisiones	Como colaborar con la IA, asegurando un equilibrio adecuado	Estudiantes, médicos
Ciencia e interpretación de datos	Comprender el análisis e interpretación de los datos médicos por parte de la IA	Estudiantes, residentes
Ética	Dilemas éticos en IA, incluyendo autonomía y privacidad del paciente	Estudiantes, formadores
Adaptación y formación continuada	Actualización continua en avances de IA a cargo de nativos en IA	Formadores, médicos

Modificado de Averbuch T, Asselbergs FW, Vardas P, Van Spall HGC. Great debate: artificial intelligence will replace much of what cardiologists do. European Heart Journal (2025) 46, 3628–3635 https://doi.org/10.1093/eurheartj/ehaf305

4 La enseñanza de la inteligencia artificial en la profesión médica

Desde el punto de vista del alumnado de grado, deberíamos contemplar dos aspectos. Uno, es la formación en IA general. Es de esperar que, al menos en un futuro próximo, los alumnos que ingresan en las facultades de Medicina posean una formación al menos básica en este campo; pero, desde luego, lo que sí que tienen es experiencia en la utilización de IA. Se trataría aquí, pues, de completar una formación reglada, pero orientada en particular a las ventajas y riesgos de la IA en Medicina.

Un principio básico es que la IA no debe entenderse como una asignatura aislada, sino como un contenido transversal, con un programa diseñado por fases.

En una primera fase (en preclínica, 1º y 2º curso) se trataría de introducir un conocimiento básico, cuyo objetivo podría definirse como alfabetización crítica. Se estructuraría en base a seminarios ("Qué es la IA y qué no", "Errores comunes al usar algoritmos") y talleres guiados ("Uso responsable de IA para estudiar", "Introducción ética temprana"). Los contenidos de estas actividades deben incluir información sobre conceptos de IA, ML (*Machine Learning*), DL (*deep learning*) y las diferencias entre modelos predictivos, modelos generativos y sistemas de apoyo a decisiones. También se debe introducir la lógica matemática sin entrar en profundidad familiarizándose con la diferencia entre correlación y causalidad y con los conceptos de probabilidad, incertidumbre y predicción, así como con el Teorema de Bayes.

Debe quedar claro que no debe aplicarse la IA a problemas médicos antes de haber cursado, y entendido, los principios de la fisiopatología, del razonamiento diagnóstico y de la incertidumbre clínica.

En el período clínico inicial (3º–4º curso) el objetivo debe ser enseñar el uso clínico supervisado, mediante el análisis de casos clínicos con y sin IA, la discusión de diagnósticos erróneos y el análisis de sesgos. Aquí deben introducirse también algunos de los aspectos de humanismo médico que comentaremos más adelante, como un taller de comunicación.

En los cursos 5º–6º debe entrarse en la práctica avanzada, con el objetivo de alcanzar una competencia profesional real. Se incluirían aquí rotaciones con herramientas de IA reales, simulaciones clínicas con IA generativa y evaluación ética de casos reales.

A lo largo de este programa, los alumnos deben aprender las utilidades indudables de la IA (Redacción de informes, resúmenes de historias clínicas, ayuda al estudio y a la docencia y simulación clínica) y sus riesgos (Alucinaciones, problemas de confidencialidad, plagio y dependencia cognitiva). Por otra parte, también deben estar informados de los posibles sesgos algorítmicos y de equidad, los problemas de transparencia y explicabilidad, la responsabilidad legal, el consentimiento informado cuando se usa IA, y el impacto en la relación médico-paciente. Las humanidades médicas son clave en estos aspectos, como veremos más adelante.

Al terminar Medicina, el estudiante debería ser capaz de:

1. Entender qué es y qué no es la IA médica
2. Usar herramientas de IA de forma crítica y segura
3. Reconocer sesgos, errores y límites
4. Explicar el uso de IA a pacientes
5. Integrar IA en el razonamiento clínico sin delegarlo
6. Actuar ética y legalmente con sistemas algorítmicos

El mensaje final de este programa es que la IA no cambia lo que significa ser médico, pero sí requiere profundizar en aquello que nunca debió perderse: el juicio, la responsabilidad y la relación humana.

Programas de este tipo ya están siendo introducidos en distintas universidades. La Universidad de Harvard implementó, a comienzos del Curso 2024-2025, un curso obligatorio de un mes para todos los estudiantes que se iniciaran en la vía HST1. Pero incluso en la vía clínica se ofertó un curso de IA en Medicina, que recibió 400 solicitudes para las siete plazas previstas. Es más, para favorecer la proactividad con respecto a la IA en Medicina, el decano de Harvard estableció los Premios a la Innovación para el uso de la IA en Educación, Investigación y Administración, dotados con

[1] La HMS (Harvard Medical School) tiene dos vías, la denominada Pathways curriculum, orientada a la clínica, que ofrece un contacto precoz y profundo con la medicina clínica, y la Harvard-MIT Health Sciences & Technology curriculum , que explora en profundidad las fronteras de la medicina y que está más orientada a investigadores y tecnólogos.

hasta 100.000 dólares para cada proyecto seleccionado. Esta iniciativa es muestra de lo que la relación entre Medicina e IA debe ser en el futuro. Los médicos y estudiantes deben formarse en IA no sólo para emplear del mejor modo algo que han desarrollado otros, sino para colaborar con ideas en el desarrollo futuro de nuevas y mejores aplicaciones a la salud.

3.2. Formación en humanismo médico.

Existe una palpable percepción, tanto entre profesionales como entre pacientes, de que la medicina está sufriendo un proceso creciente de deshumanización. Este proceso no es todavía atribuible al advenimiento de la IA, sino que posiblemente obedece a diversas causas que no es el momento de analizar aquí en profundidad. Pero entre ellas podemos citar una tecnificación y encorsetamiento del proceso de manejo de las distintas enfermedades, de la mano de las Guías de Práctica Clínica y de las Vías Clínicas, que equiparan el proceso de la enfermedad a los procedimientos industriales en serie. No cabe duda de que este tipo de medicina es mucho más eficiente a la hora de hacer frente a la creciente demanda de la población, y que garantiza las buenas prácticas clínicas. Este abordaje hace necesaria, al menos en enfermedades crónicas, la actuación de lo que se ha dado en denominar equipos multidisciplinares. También esto es positivo, pues favorece la coordinación entre expertos en los distintos aspectos de la enfermedad. El inconveniente es que difumina la relación interpersonal médico-paciente. A éste le resulta difícil la mayor parte de las veces identificar a "su médico" con un profesional concreto, con nombre y cara. Por otra parte, el envejecimiento de la población en un escenario en que la demanda en salud es ilimitada, pero los recursos para atenderla no, hace que cada vez se disponga de menos tiempo para cada enfermo, por lo que los contactos deben limitarse a tratar los aspectos puramente clínicos de la enfermedad (síntomas), obviando el contexto, o la repercusión que la percepción de la enfermedad tiene sobre el paciente (dolor, sufrimiento, desesperación o miedo a la muerte). La medicina defensiva, que a algunos médicos les hace sentir que los enfermos son potenciales enemigos, no facilita las cosas. Y la historia clínica electrónica tampoco, pues centra la atención del médico en la pantalla del ordenador en vez de mirar a la cara a su enfermo.

La aplicación de ciertas herramientas derivadas de la IA, como los AMS (Ambient Medical Scribes), capaces de registrar la entrevista clínica y redactarla en forma de nota clínica, es una de las formas en que la IA podría ayudar a mejorar la percepción de una medicina más humana y a reducir

el burnout de los médicos. Aunque los logros objetivables de este tipo de herramientas, ya disponibles en muchos hospitales de nuestro país, son aún modestos, es de esperar que su perfeccionamiento los haga aún más efectivos. Una de sus ventajas es que reduce el tiempo de consulta; pero esto puede conducir no a disponer de más tiempo para cada paciente, sino a la exigencia de que un mismo médico atienda a más pacientes. También la capacidad de traducción instantánea que posibilita la IA entre múltiples lenguas puede favorecer la comunicación en un mundo pluricultural.

La necesidad, pues, de introducir acciones dentro del pregrado que conduzcan a una práctica más humana de la medicina, antecede al advenimiento de la IA en la práctica clínica. La formación humanística contribuye, además, a reducir el burnout, a la mejora personal de los médicos y a desarrollar una mayor empatía. Numerosas iniciativas han surgido a nivel internacional, como el Hippocratic Movement, o los Capítulos Hipocráticos dentro de sociedades científicas, como la Sociedad Española de Cardiología.

También existe un creciente interés por incorporar contenido humanístico a la docencia. Por ejemplo, la universidad Johns Hopkins ofrece cursos de Medicina Narrativa (MN), con talleres multidisciplinarios de escritura reflexiva, arte y literatura para promover la empatía y la comprensión de la experiencia del paciente.

La IA obligará aún más a profundizar en este tipo de formación. Una idea adecuada en este sentido es ligar la formación humanística al entorno que supone la IA. En la facultad de Medicina de la Universidad de Málaga existen, desde hace dos años, una serie de seminarios sobre IA y Humanismo Médico, como iniciativa individual dentro de la Cátedra de Cardiología.

Es probable que en un futuro muy próximo los distintos modelos de IA puedan hacerse cargo de una serie de tareas en la asistencia médica, complementando o substituyendo la labor del médico. No parece probable, al menos a corto plazo, que la figura del médico se vea amenazada; pero sí es conveniente reconducir su preparación para reforzar aquellos aspectos en que la IA no puede reemplazarle. Uno de estos aspectos es la empatía. Los modelos de IA generativa basados en el machine learning entrenados en datos multimodales pueden analizar claves vocales, faciales y posturales para proporcionar al médico información del sentir de los pacientes. Otra cosa muy distinta es la vía inversa. Es cierto que avatares diseñados con IA pueden simular empatía verbal; pero esto es solo un componente de

la empatía. El lenguaje no verbal y el contacto físico son partes esenciales para entender y compartir los sentimientos del otro. Lewis Thomas, en su primera clase a estudiantes de medicina solía decir, en tono provocador, que, si pudiera, "rociaría la sala con el virus de la gripe" para que aquellos que nunca habían estado realmente enfermos pudieran experimentar lo que eso significa, pues "Para ser médico", seguía, "necesitas empatía; para eso, tienes que haber experimentado la enfermedad". Solo quien ha estado enfermo, o ha vivido de cerca la enfermedad, puede entender lo que experimenta el paciente. Del mismo modo, solo el que sabe que es mortal puede entender al enfermo cuando se enfrenta a la muerte. Estos sentimientos son patrimonio del ser humano; estos son los terrenos donde el médico no puede ser substituido.

Por otra parte, entender la enfermedad no solo como un proceso patológico, sino como una experiencia vivencial que afecta y se ve afectada por el entorno socioeconómico y cultural del enfermo, por sus creencias, miedos y expectativas, requiere una formación, al menos básica, en muchos otros aspectos que aparentemente no guardan relación con la medicina. Tales son la literatura, filosofía, historia, etc. Es cierto que este tipo de formación debería recibirla el estudiante de medicina durante la enseñanza secundaria, aunque los sucesivos planes de este nivel de enseñanza van primando cada vez más la formación técnica sobre la humanística, entendiendo por esta última todo lo no directamente utilitario. Por ello, se deben contemplar acciones para favorecer este tipo de formación durante los años de grado, e incentivar el que los alumnos sigan cultivando estas disciplinas en el postgrado. En la Universidad de Málaga funciona desde hace años el Club Hipócrates, en que de forma periódica estudiantes de Medicina y profesores se reúnen para hablar sobre cualquier tema, excepto de Medicina. Es esta una iniciativa no oficial ni reglada. Pero ya existen universidades donde se han incorporado iniciativas formales en este sentido. Hemos mencionado la universidad Johns Hopkins y su enseñanza de la MN. Rita Charon, que dirige el Programa de Medicina Narrativa de la Universidad de Columbia, define la MN como el desarrollo de habilidades para reconocer, absorber, interpretar y sentirse conmovido por las historias que rodean a la enfermedad, que puede emplearse para entrenar a los profesionales de la salud para ser capaces de contemplar el sufrimiento de los pacientes. En contraste con la medicina basada en la evidencia, que se centra en la enfermedad y en los aspectos científicos de la medicina, la MN prioriza sus aspectos éticos y humanos, centrándose en la comunicación entre los enfermos, sus familias y los profesionales sanitarios, para lo cual son fundamentales las habilidades

lingüísticas. Las metáforas, la ironía y la gestualidad son aspectos difícilmente integrables en los algoritmos de la IA. Precisamente la Universidad de Columbia, en Nueva York, incorpora desde 2014 como obligatoria una asignatura sobre MN en su curriculum. Otras universidades, como Harvard, Georgetown, Florida...ofrecen esta asignatura como optativa. En Europa, la Universidad del Sur de Dinamarca la incorpora como obligatoria, mientras que la Complutense de Madrid, la Universidad de Lisboa o la de Utrecht la ofrecen como optativa. Hay datos que sugieren que el entrenamiento en MN mejora una serie de aspectos en los que lo reciben, incluyendo la catarsis emocional y las habilidades de expresión escrita y hablada, aunque aún no se dispone de información sobre su impacto en la percepción de los pacientes o el desarrollo final de la enfermedad.

No es la única herramienta diseñada para favorecer la formación humanística en la medicina. La Asociación Europea de Educación Médica (AMEE) recomienda, en su Guía 179, la implementación de *Visual Thinking Strategies* (Estrategias de Pensamiento Visual, VTS). Al comienzo del documento, se afirma de forma expresa que "A medida que la IA influye cada vez más sobre el panorama de la asistencia médica, se hace necesario integrar las humanidades médicas en la educación sanitaria para cultivar el pensamiento crítico y la empatía en el cuidado de los pacientes" La propuesta de la AMEE, el VTS, fue desarrollado inicialmente por Housen y Yenawine en 2013, basándose en una experiencia que la Universidad de Harvard inició en 2004, "*Training the Eye: Improving the Art of Physical Diagnosis*" (Entrenando el ojo: mejorando el Arte del Diagnóstico Físico"). Ese tipo de iniciativas ha encontrado eco; una revisión de 2022 reveló que existen al menos 125 programas basados en la colaboración entre museos de arte y facultades de Medicina.

Los objetivos de la educación en humanidades médicas dentro de las facultades de medicina se han ido formulando con bastante consenso internacional (EE. UU., Reino Unido y Europa), especialmente a partir de la bioética, la medicina narrativa y las humanidades médicas. No se trata de "hacer médicos más cultos", sino de formar mejores clínicos, profesionales y ciudadanos.

Los objetivos suelen agruparse en varios puntos:

1. Comprender la enfermedad como experiencia humana

Objetivo: que el estudiante entienda que la enfermedad no es solo un proceso biológico, sino una experiencia vivida.

* Distinguir enfermedad biológica de experiencia subjetiva del enfermar.

* Comprender el impacto de la enfermedad en la identidad, la biografía, la familia y el entorno social del paciente.

* Reconocer el sufrimiento, la vulnerabilidad y la incertidumbre como partes constitutivas de la práctica médica.

Aquí se inscriben la medicina narrativa, y la antropología médica.

2. Desarrollar empatía clínica y capacidad de escucha

Objetivo: formar médicos capaces de escuchar, interpretar y responder a las historias de los pacientes.

* Mejorar la escucha activa y la atención al relato del paciente.

* Reconocer silencios, metáforas, emociones y contradicciones en la narrativa clínica.

* Evitar la deshumanización del paciente como “caso” o “diagnóstico”.

Importante: las humanidades no buscan sentimentalismo, sino empatía profesional, compatible con el rigor clínico.

3. Mejorar la comunicación clínica

Objetivo: formar médicos que se comuniquen mejor con pacientes, familias y equipos sanitarios.

* Comunicar malas noticias con sensibilidad.

* Explicar diagnósticos y tratamientos de forma comprensible.

* Manejar conflictos, expectativas irreales y situaciones de alta carga emocional.

* Fomentar la toma de decisiones compartida.

Este objetivo está estrechamente ligado a la literatura, la narrativa y el análisis del lenguaje.

4. Formar el juicio ético y la responsabilidad profesional

Objetivo: ayudar al estudiante a pensar éticamente, no solo a aplicar normas.

* Analizar dilemas morales reales (final de la vida, consentimiento, justicia distributiva).

* Reconocer conflictos de valores (del paciente, del médico, de la institución).

* Comprender el papel social del médico y su responsabilidad pública.

* Integrar bioética, filosofía moral e historia de la medicina.

Las humanidades enseñan a deliberar, no a memorizar códigos.

5. Fomentar la reflexión crítica sobre la medicina

Objetivo: que el futuro médico sea capaz de pensar críticamente su propia práctica.

* Cuestionar el tecnicismo excesivo y el reduccionismo biomédico.

* Reflexionar sobre los límites de la medicina.

* Entender la medicina como práctica histórica, cultural y socialmente situada.

* Analizar el poder médico y la relación asimétrica médico-paciente.

Aquí entran la historia de la medicina, la sociología, la filosofía y los estudios culturales.

6. Prevenir el burnout y fortalecer la identidad profesional

Objetivo: cuidar al médico en formación.

* Ofrecer espacios de escritura reflexiva y discusión para procesar experiencias difíciles.

* Ayudar a construir una identidad profesional coherente y consciente.

* Favorecer el autocuidado, la resiliencia y el sentido del trabajo médico.

* Reducir la despersonalización y el cinismo.

Muchas facultades introducen las humanidades como respuesta al burnout en estudiantes y residentes.

7. Promover sensibilidad cultural y la justicia social

Objetivo: formar médicos capaces de atender a pacientes diversos en contextos complejos.

* Comprender cómo la cultura, el género, la clase social o la migración influyen en la salud.

* Reconocer desigualdades estructurales y determinantes sociales de la enfermedad.

* Evitar estereotipos y sesgos implícitos.

* Promover una medicina más equitativa y socialmente responsable.

8. Complementar (no sustituir) la ciencia biomédica

Objetivo clave: integrar humanidades y ciencia, no enfrentarlas.

* Las humanidades no reemplazan la biología, la fisiología o la evidencia científica.

* Ayudan a aplicar mejor ese conocimiento en situaciones reales, humanas e inciertas.

* Forman médicos técnicamente competentes y humanamente atentos.

En resumen, la educación en humanidades médicas busca formar médicos que: comprendan la enfermedad como experiencia humana, escuchen y se comuniquen mejor, actúen con empatía y ética, reflexionen críticamente sobre su práctica, cuiden de sí mismos y de los demás e integren ciencia, técnica y humanidad.

3.3. Modificaciones en la enseñanza del grado de Medicina para adaptarse a un mundo cambiante

Está claro que de todo lo anterior se sigue que deben introducirse nuevos contenidos en la enseñanza de las Facultades de Medicina. Los ya sobrecargados contenidos de los actuales planes de estudio hacen que el añadir

contenidos nuevos solo pueda hacerse a expensas de reducir otros, cuya utilidad en el futuro es dudosa. La labor de decidir qué contenidos suprimir se plantea plagada de dificultades. En el sistema universitario español es de prever una lucha sin cuartel entre los distintos Departamentos por mantener su carga docente, cuyo tamaño justifica el numero de puestos docentes, que es visto como base de poder. En parte el problema viene agravado por la diferente representación cuantitativa en las plantillas de profesorado permanente a favor de los profesores de ciencias básicas en detrimento de los clínicos. La política de acreditaciones de la ANECA (Agencia Nacional de Evaluación y Acreditación española) no es ajena a este desequilibrio, al primar a aquéllos que disponen de más tiempo para la investigación sobre los que se ven constreñidos por las exigencias de la asistencia clínica. Pero es preciso alcanzar acuerdos pronto, pues de otro modo el lapso entre las universidades con capacidad de adaptarse a las situaciones cambiantes de forma ágil, y aquéllas otras ancladas en la tradición y en procedimientos burocráticos y tediosos, no hará sino agrandarse.

Por otra parte, la IA puede ayudar a paliar alguno de los problemas que arrastran muchas de nuestra universidades. La desproporción entre el número de alumnos y el de recursos asistenciales para hacer prácticas podría parcialmente compensarse mediante la simulación clínica y la aplicación de sistemas de enseñanza tutorizada basada en modelos socráticos. Pero esto tampoco puede depender en exclusiva de iniciativas individuales, sino que debe ser política, al menos, de cada institución universitaria.

4. Formacion en IA durante la residencia

En un futuro próximo es seguro que los alumnos que se licencien en las Facultades de Medicina lo harán con una adecuada formación que les permita emplear de forma correcta las herramientas que la IA pone a su disposición, sobre todo si se articula e implementa un programa reglado de formación en el sentido anteriormente desarrollado. Pero no es esa la situación ahora. Vivimos en una época de cambio permanente en que muchos médicos emplean la IA como una ayuda en sus tareas clínicas, sin ser conscientes de los riesgos y limitaciones que ello comporta. Algunos de estos riesgos son bien conocidos:

1. Las alucinaciones, que suceden cuando un LLM hace afirmaciones taxativas, pero falsas, que pueden ser tomadas por verdaderas. Es cierto que los modelos más avanzados alucinan menos que los iniciales; pero también lo es que el problema se ve agravado por la poca transparencia del proceso por el que los modelos de IA alcanzan sus conclusiones (*black box*). Además, las alucinaciones pueden adoptar formas más difíciles de detectar, como inventarse citas falsas, incluso de revistas científicas que no existen. El Comité Internacional de la Cruz Roja ha advertido recientemente sobre este hecho, recomendando que para detectar el error se consulte con los catálogos oficiales de publicaciones científicas.

2. La pérdida de habilidades (*deskilling*) que puede ser consecuencia de utilizar con demasiada frecuencia la IA. Algunos estudios han objetivado este hecho: Endoscopistas entrenados pueden ver mermadas sus habilidades diagnósticas tras tres meses de ayudarse de un modelo de IA basado en el reconocimiento de imágenes para detectar pólipos en el colon. Del mismo modo, se ha demostrado que existe una asociación negativa significativa entre el uso frecuente de la IA y la capacidad de pensamiento crítico. En el estudio que analizó este hecho se demostró que los participantes más jóvenes presentaban una mayor dependencia de las herramientas de IA y un peor pensamiento crítico que los participantes de más edad, lo que sugiere que el abuso de IA puede dificultar el desarrollo de este tipo de pensamiento.

3. En el mismo sentido, es aún peor la no adquisición de habilidades (*never skilling*) por confiar y emplear en demasía la IA durante el periodo de formación.

4. La capacitación inadecuada (*miskilling*) que sucede cuando se confía en exceso en modelos de IA alimentados y entrenados con sesgos de raza, género u otro tipo, lo que puede conducir a la perpetuación de esos mismos sesgos en los médicos afectados.

5. El exceso de confianza (*over-reliance*) y su consecuencia, el sesgo de automatización (*automation bias*) consistente en confiar, sin análisis crítico, en las recomendaciones diagnósticas y terapéuticas de los modelos de la IA, lo que, aparte de propiciar errores y perpetuar sesgos, puede agravar todo lo anterior.

4 La enseñanza de la inteligencia artificial en la profesión médica

Todos estos riesgos están ya presentes en el día a día del quehacer clínico. En un reciente artículo de revisión, Abdulnour y cols. plantean el siguiente escenario, habitual ya en nuestros hospitales: Durante una sesión clínica, un residente consulta su *smartphone* tras evaluar a un paciente y obtiene en segundos, tras introducir el *prompt* adecuado en un LLM, un bien argumentado y completo plan de diagnóstico diferencial y plan terapéutico. El tutor clínico se plantea una serie de preguntas: ¿Cuál ha sido el *prompt* introducido?¿Cuestiona el residente la respuesta de la IA, o la acepta sin análisis crítico?¿Deberíamos fiarnos de esa respuesta?¿Debo intervenir o no? Y, por último ¿Es este el futuro del razonamiento clínico? Los autores, de las Universidades de Harvard, Illinois y California-San Francisco, señalan algo que ya hemos comentado: la probabilidad de que el médico más joven -el residente- esté más familiarizado con la IA que su tutor, de más edad. La estrategia que proponen para enfrentar estas situaciones y entrenar a los residentes en la utilización adecuada de las herramientas que la IA nos ofrece incluye, por supuesto, la necesidad de formar a los formadores. Pero además la adopción de una estrategia que ellos denominan DEFT-AI para promover el pensamiento crítico. DEFT (*Diagnosis, Evidence, Feedback, and Teaching* o Diagnóstico, Pruebas, Retroalimentación y Enseñanza) es una estrategia ya existente que los autores proponen adaptar al entorno de la IA en la clínica. Aplicándola al ejemplo anterior, el tutor debe comenzar por preguntar al residente su razonamiento clínico, qué *prompt* empleó y si la respuesta modificó su razonamiento previo. A continuación, siempre siguiendo un método socrático, basado en preguntas y respuestas, el tutor inicia la discusión sobre el diagnóstico diferencial, en qué se basa -datos que apoyan o van en contra de cada posibilidad- qué evidencia hay al respecto y cuál es la precepción del educando sobre la IA en este contexto. La retroalimentación es precedida por una autoevaluación por parte del residente. Por último, la fase de enseñanza incluye no solo la parte clínica sino también las posibilidades y riesgos de la IA.

Es probable que la colaboración médico-IA adopte una de dos formas, que Abdulnour y cols. denominan *centauro* o *ciborg*. En la primera, los médicos dividen las tareas entre las propias y las que delegan en la IA, reservando para sí las que atañen al juicio clínico y la toma de decisiones. Esta estrategia es la recomendable para casos complejos, inciertos o de alto riesgo. La forma ciborg supone una estrecha colaboración, sin límites claros o preestablecidos, entre hombre e IA. Aunque esta estrategia puede ser eficiente en casos de bajo riesgo o bien estandarizados, implica los

riesgos que hemos comentado anteriormente de pérdida de habilidades o de exceso de confianza.

Este mismo proceso de formación puede ser aplicado también en un entorno diferente, fuera ya de los ámbitos académicos o del período de formación hospitalaria, que plantea problemas específicos: el de los médicos en ejercicio extrahospitalarios o que trabajan en centros hospitalarios que no han implementado programas de formación en este campo.

5. Papel de los Colegios de Médicos

El colectivo descrito en el ultimo párrafo es particularmente vulnerable. Están expuestos al entorno de la IA y compelidos a emplearla, pero para hacerlo con seguridad y eficacia necesitan la adecuada formación. Es cierto que existe una amplia, pero desigual, oferta de formación por parte de instituciones varias, públicas y privadas. Pero se hace necesaria la implantación de un programa adecuado, que incluya muchos de los contenidos que hemos mencionado anteriormente, no descuidando los componentes éticos y humanísticos. El establecimiento de un programa de este tipo es un reto y una oportunidad para instituciones como los Colegios de Médicos, que ofrecerían así a sus colegiados un apoyo que muchos sienten que necesitan, al mismo tiempo que garantizarían, con la autoridad que ostentan, el que la formación sea la adecuada.

La puesta en marcha de estas iniciativas debería tener en cuenta los siguientes puntos:

1. Un programa nacional acreditado, que debería surgir de la iniciativa del Consejo General, que éste pondría a disposición de todos los Colegios.

2. Un adecuado incentivo para los potenciales alumnos. Este podría ser positivo, en forma de créditos de formación continuada u otras formas que puedan arbitrarse; o negativo. La obligatoriedad de conseguir una capacitación mínima en el uso de la IA en Medicina es posible en aquellos países en que existe la recertificación periódica.

3. Deben habilitarse seminarios periódicos de actualización para aquellos que hayan hecho el curso de formación de, no lo olvidemos, un campo que evoluciona día a día.

6. Evaluación y acreditación de competencias en IA

La incorporación de la IA a la enseñanza médica no puede limitarse a la transmisión de conocimientos ni al entrenamiento informal en el uso de herramientas. Para que la formación sea efectiva, segura y responsable, es imprescindible definir, evaluar y acreditar competencias específicas en IA médica. En ausencia de sistemas de evaluación claros, existe el riesgo de que la IA se convierta en un recurso utilizado de forma acrítica, desigual y potencialmente peligrosa.

La evaluación de competencias en IA plantea desafíos particulares. A diferencia de otras habilidades clínicas, no se trata de medir la capacidad de obtener una respuesta correcta, sino de valorar el proceso de razonamiento, el juicio crítico, la comprensión de los límites del sistema y la integración ética de la herramienta en la práctica clínica. En este sentido, evaluar el "uso adecuado" de la IA es más relevante que evaluar el "resultado" proporcionado por la IA.

La evaluación debería estructurarse en torno a varios dominios claramente definidos:

- Comprensión conceptual

El estudiante o médico debe demostrar que entiende qué es y qué no es la IA médica, sus principios básicos, diferencias entre modelos predictivos y generativos, y los conceptos de incertidumbre, sesgo y explicabilidad.

- Razonamiento clínico asistido por IA

Capacidad para integrar la información proporcionada por sistemas de IA sin delegar el juicio clínico, identificando concordancias y discrepancias con su propio razonamiento y justificando decisiones diagnósticas y terapéuticas.

- Pensamiento crítico y detección de errores

Habilidad para reconocer alucinaciones, inconsistencias, recomendaciones inadecuadas o falta de evidencia, así como para cuestionar activamente las respuestas de los sistemas algorítmicos.

- Uso ético, legal y profesional

Conocimiento de los principios de confidencialidad, consentimiento informado, responsabilidad profesional y límites del uso de la IA, así como de los riesgos de dependencia cognitiva y sesgo de automatización.

- Comunicación clínica

Capacidad para explicar al paciente, de forma comprensible y honesta, el papel que ha desempeñado la IA en su proceso diagnóstico o terapéutico, manteniendo la confianza y la relación médico-paciente.

6.1. Métodos de evaluación propuestos

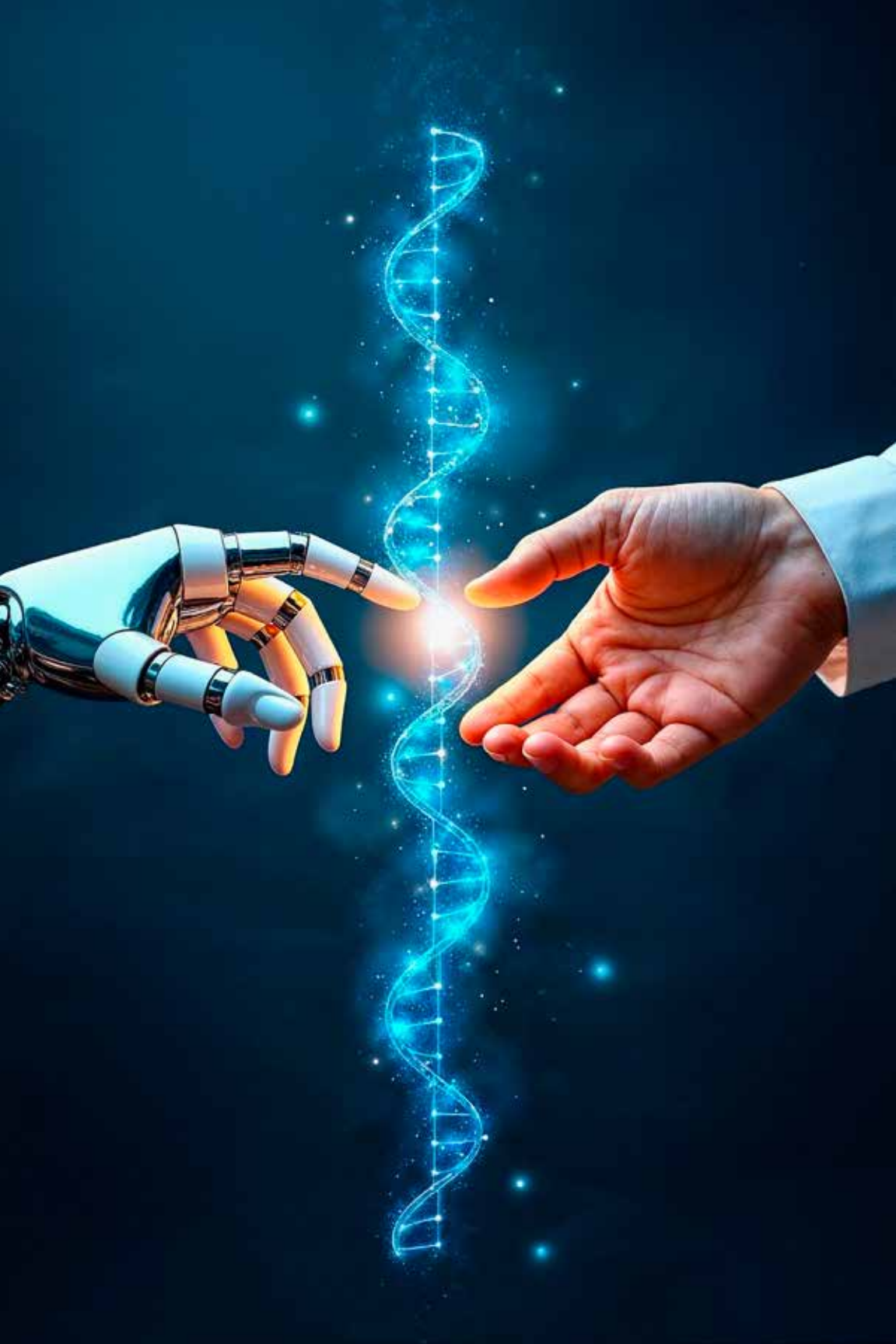
1. ECOEs con integración explícita de IA

Las Evaluaciones Clínicas Objetivas Estructuradas (ECOE) constituyen un entorno idóneo para evaluar competencias en IA médica. En este contexto, el alumno puede disponer de una herramienta de IA durante la estación clínica, y la evaluación debe centrarse en cómo formula la consulta a la IA (*prompting*), cómo interpreta la respuesta, si identifica errores o limitaciones, y cómo integra (o rechaza) la información en su decisión final.

El objetivo no es penalizar el uso de la IA, sino evaluar la calidad del razonamiento clínico aumentado.

2. Análisis crítico de casos con IA

Se pueden presentar casos clínicos resueltos parcialmente con ayuda de IA, incluyendo ejemplos con errores deliberados, sesgos o recomendaciones



inapropiadas. El estudiante debe identificar los problemas, justificar por qué la recomendación es errónea o incompleta y proponer una alternativa razonada.

Este método es especialmente útil para evaluar pensamiento crítico y evitar el sesgo de automatización, y de hecho ya se ha incorporado en determinados programas docentes. Por ejemplo, se ha implantado en el curso académico 2025-2026 en la Asignatura de Enfermedades del Aparato Circulatorio de la Facultad de Medicina de la Universidad de Málaga. El análisis de los resultados revela una elevada valoración por parte de los estudiantes, destacando que ayuda a conocer las fortalezas y debilidades de aplicaciones de IA en el "mundo real"

3. Evaluación reflexiva estructurada

El uso de ensayos cortos de reflexión permite evaluar aspectos que no son fácilmente cuantificables, como la percepción del propio aprendizaje, la conciencia de los límites de la IA, y el impacto del uso de IA en la identidad profesional del estudiante.

Estos instrumentos son particularmente valiosos en la prevención del *deskilling* y del *never skilling*.

4. Evaluación del razonamiento, no del resultado

En pruebas escritas u orales, se debe pedir explícitamente al alumno que describa su razonamiento previo al uso de la IA, explique cómo la respuesta de la IA influyó (o no) en su decisión y justifique por qué aceptó o rechazó las recomendaciones.

Esto refuerza la idea de que la IA es una herramienta de apoyo, no una autoridad.

5. Evaluación en el entorno clínico real (*workplace-based assessment*)

Durante las rotaciones clínicas y la residencia, los tutores pueden evaluar competencias en IA mediante observación directa del uso de IA en casos reales, discusión estructurada posterior al caso y herramientas tipo mini-CEX (*Clinical Evaluation Exercise*) adaptadas al contexto de la IA.

Este enfoque es coherente con estrategias como el modelo DEFT-AI y permite integrar la evaluación en la práctica cotidiana.

6.2. Acreditación de competencias

La adquisición de competencias en IA debería culminar en algún tipo de acreditación formal, especialmente en el postgrado y la formación continuada. Esta acreditación podría adoptar varias formas, como la certificación universitaria al finalizar el grado, el reconocimiento como competencia transversal en el título de especialista, los créditos de formación continuada acreditados por colegios profesionales o los programas de recertificación periódica en sistemas sanitarios que los contemplen.

Más que certificar el dominio de herramientas concretas —que cambian rápidamente—, estas acreditaciones deben centrarse en competencias transferibles: juicio clínico, pensamiento crítico, ética y comunicación en un entorno mediado por IA.

7. Conclusiones

Es preciso incorporar la formación en IA dentro de la enseñanza de la Medicina, tanto a nivel de pregrado como de postgrado. Al mismo tiempo debe reforzarse la formación en aspectos de los que se denomina medicina humanística. El desarrollo de metodologías y programas adecuados, junto con la estructuración de los correspondientes sistemas de evaluación y acreditación, son tareas urgentes que deben ser emprendidas con premura por las instituciones académicas y profesionales.

8. Bibliografia

1. Citado por Khan, Salman. Brave New Words: How AI Will Revolutionize Education (and Why That's a Good Thing) (English Edition) (p. 4). (Function). Kindle Edition.

2. https://www.khanacademy.org

3. https://magazine.hms.harvard.edu/articles/how-generative-ai-transforming-medical-education

4. Elias P, Black KC, Chandak, P et al. The Moment AI Arrived in the Clinic: Insights from the SAIL 2025 Year in Review. NEJM AI 2025;2(11) DOI: 10.1056/AIp2500788

5. Salman Khan BRAVE NEW WORDS How AI Will Revolutionize Education and Why That's a Good Thing). Penguin Random House UK, Londres 2024. Kindle Edition.

6. https://arxiv.org/abs/2512.05671

7. Zhitao He, Haolin Yang, Zeyu Qin, Yi R Fung. MedTutor-R1: Socratic Personalized Medical Teaching with Multi-Agent Simulation. https://chatpaper.com/es/paper/216549

8. https://time.com/6306922/artificial-intelligence-medicine-doctors/

9. Averbuch T, Asselbergs FW, Vardas P, Van Spall HGC. Great debate: artificial intelligence will replace much of what cardiologists do. European Heart Journal (2025) 46, 3628–3635 https://doi.org/10.1093/eurheartj/ehaf305

10. https://magazine.hms.harvard.edu/articles/how-generative-ai-transforming-medical-education

11. Kim E, Liu VX, Singh K. AI Scribes Are Not Productivity Tools (Yet) NEJM AI 2025;2(12) DOI: 10.1056/AIe2501051

12. Misurac J, Knake LA, Blum JM. The effect of ambient artificial intel¬ligence notes on provider burnout. Appl Clin Inform 2025;16:252- 258. DOI: 10 .1055 /a -2461 -4576.

13. Mangione S, Chakraborti C, Staltari G, Harrison R, Tunkel AR, Liou KT, et al. Medical Students' Exposure to the Humanities Correlates with Positive Personal Qualities and Reduced Burnout: A Multi-Institutional U.S. Survey. Journal of General Internal Medicine. 2018;33(5):628–34.

14. Graham J, Benson LM, Swanson J, Potyk D, Daratha K, Roberts K. Medical Humanities Coursework Is Associated with Greater Measured Empathy in Medical Students. Am J Med. 2016;129(12):1334–7.

15. https://hippocratic-movement.org

16. https://secardiologia.es/institucional/reuniones-institucionales/capitulo-hipocratico-de-la-sec

17. Mazumdar, Showvik. 2024. The role of storytelling in human advancement: A reflection on technology, empathy, and learning. https://www.bciit.ac.in/pdf/E-Magazine/Finalmergedemagazine2024.pdf#page=10

18. Thomas, L. Las vidas de una célula: notas de un observador de la biología, 1974, Viking Press: Penguin Books, reimpresión de 1995:ISBN 0-14-004743-3

19. Charon R. Narrative Medicine: honoring the Stories of Illness. Oxford, UK: Oxford University Press; 2006.

20. Hung-Chang Liao & Ya-Huei Wang (2023) Narrative medicine and humanities for health professions education: an experimental study, Medical Education Online, 28:1, 2235749, DOI: 10.1080/10872981.2023.2235749

21. M. M. Milota, G. J. M. W. van Thiel & J. J. M. van Delden (2019) Narrative medicine as a medical education tool: A systematic review, Medical Teacher, 41:7, 802-810, DOI: 10.1080/0142159X.2019.1584274

22. Agarwal, G., Yenawine, P., Manohar, S., & Chisolm, M. S. (2025). Implementing a Visual Thinking Strategies program in health professions schools: An AMEE Guide for health professions educators: AMEE Guide No. 179. Medical Teacher, 47(9), 1425–1434. https://doi.org/10.1080/0142159X.2025.2458287

23. Cerqueira, A.R., Alves, A.S., Monteiro-Soares, M. et al. Visual Thinking Strategies in medical education: a systematic review. BMC Med Educ 23, 536 (2023). https://doi.org/10.1186/s12909-023-04470-3

24. Yenawine P. Visual Thinking Strategies: Using Art to Deepen Learning Across School Disciplines. Cambridge, MA: Harvard Education Press; 2013.

25. Harvard Medical School- Training our Eyes, Minds and Hearts: Visual Thinking Strategies for Health Care Professionals. https://cmecatalog.hms.harvard.edu/training-our-eyes-minds-and-hearts-visual-thinking-strategies-health-care-professionals.

26. Pitman B. Art Museum and Medical School Partnerships: Program Descriptions. In: The Edith O'Donnell Institute of Art History. The University of Texas at Dallas. 2022. https://arthistory.utdallas.edu/medicine/resources/2022%20PROGRAM%20DESCRIPTIONS_FINAL.pdf

27. Vergano D. AI Slop Is Spurring Record Requests for ImaginaryJournals. Scientific American, Diciembre 2025, https://www.scientificamerican.com/article/ai-slop-is-spurring-record-requests-for-imaginary-journals/

28. https://www.icrc.org/en/article/important-notice-ai-generated-archival-references

29. Budzyń, Krzysztof et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. The Lancet Gastroenterology & Hepatology, Volume 10, Issue 10, 896 – 903

30. Gerlich M. AI tools in society: impacts on cognitive offloading and the future of critical thinking. Societies (Basel) 2025; 15: 6 (https://www.mdpi.com/2075-4698/15/1/6).

31. Rafel JB. Proceedings and abstracts of the 2024 Artificial Intelligence and Medical Education Macy Conference, November 18, 2024. Atlanta: Josiah Macy Jr. Foundation, 2024.

32. Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA 2023; 330: 2275-84.

33. Hough J, Culley N, Erganian C, et al. Potential risks of GenAI on medical education. BMJ Evidence-Based Medicine 2025;30:406-408.

34. Abdulnour RAE, Gin B, Boscardin CK. Educational Strategies for Clinical Supervision of Artificial Intelligence Use. N Engl J Med 2025;393:786-97. DOI: 10.1056/NEJMra2503232

35. Savaria MC, Min S, Aghagoli G, Tunkel AR, Hirsh DA, Michelow IC. Enhancing the one-minute preceptor method for clinical teaching with a DEFT approach. Int J Infect Dis 2022; 115: 149-53

5

Profesionalismo médico en el contexto de la Inteligencia Artificial

Dr. Francisco Miralles Linares
Jefe de Servicio de Medicina Interna
Hospital Vithas Xanit Internacional

Resumen ejecutivo

La irrupción de la inteligencia artificial (IA) en la medicina genera un desafío en los fundamentos del profesionalismo médico tradicional. La cuestión radica en cómo integrar herramientas algorítmicas potentes sin erosionar la autonomía, la ética y la confianza inherentes a la relación médico-paciente (1,2). Su relevancia es enorme. La IA promete mejorar la precisión diagnóstica y aliviar cargas burocráticas (3), pero también podría introducir riesgos de deshumanización, sesgos y difuminación de responsabilidades (8,9). La IA puede actuar como copiloto del médico, aumentando su capacidad cognitiva y agilizando decisiones clínicas, pero no debería reemplazar el juicio clínico ni la responsabilidad final del facultativo (6,13). La incorporación acrítica de algoritmos opacos podría minar la autonomía profesional y la decisión terapéutica, mientras que un uso reflexivo y ético puede potenciar la calidad de la atención. Las implicaciones para el profesionalismo médico son claras. Es necesario redefinir el rol del médico como líder humanista en la era digital, capaz de supervisar la IA, mantener la empatía y asumir nuevas responsabilidades compartidas sin renunciar a los valores tradicionales de la profesión.

Palabras clave:

Profesionalismo médico; inteligencia aumentada; autonomía profesional; sesgo de automatización; explicabilidad y transparencia algorítmica; trazabilidad y reversibilidad; rendición de cuentas y responsabilidad; relación médico-paciente; empatía y humanismo.

Executive summary:

The emergence of artificial intelligence (AI) in medicine poses a challenge to the foundations of traditional medical professionalism. The key issue is how to integrate powerful algorithmic tools without eroding the autonomy, ethics, and trust inherent to the physician–patient relationship (1,2). Its relevance is substantial. AI promises to improve diagnostic accuracy and reduce bureaucratic burdens (3), but it could also introduce risks of dehumanization, bias, and a blurring of responsibilities (8,9). AI can function as

a physician's copilot, enhancing cognitive capacity and streamlining clinical decision-making, but it should not replace clinical judgment or the clinician's ultimate responsibility (6,13). The uncritical adoption of opaque algorithms could undermine professional autonomy and therapeutic decision-making, whereas a reflective and ethical use can enhance the quality of care. The implications for medical professionalism are clear. The physician's role must be redefined as that of a humanistic leader in the digital era—capable of supervising AI, preserving empathy, and assuming new shared responsibilities without relinquishing the profession's traditional values.

Keywords:

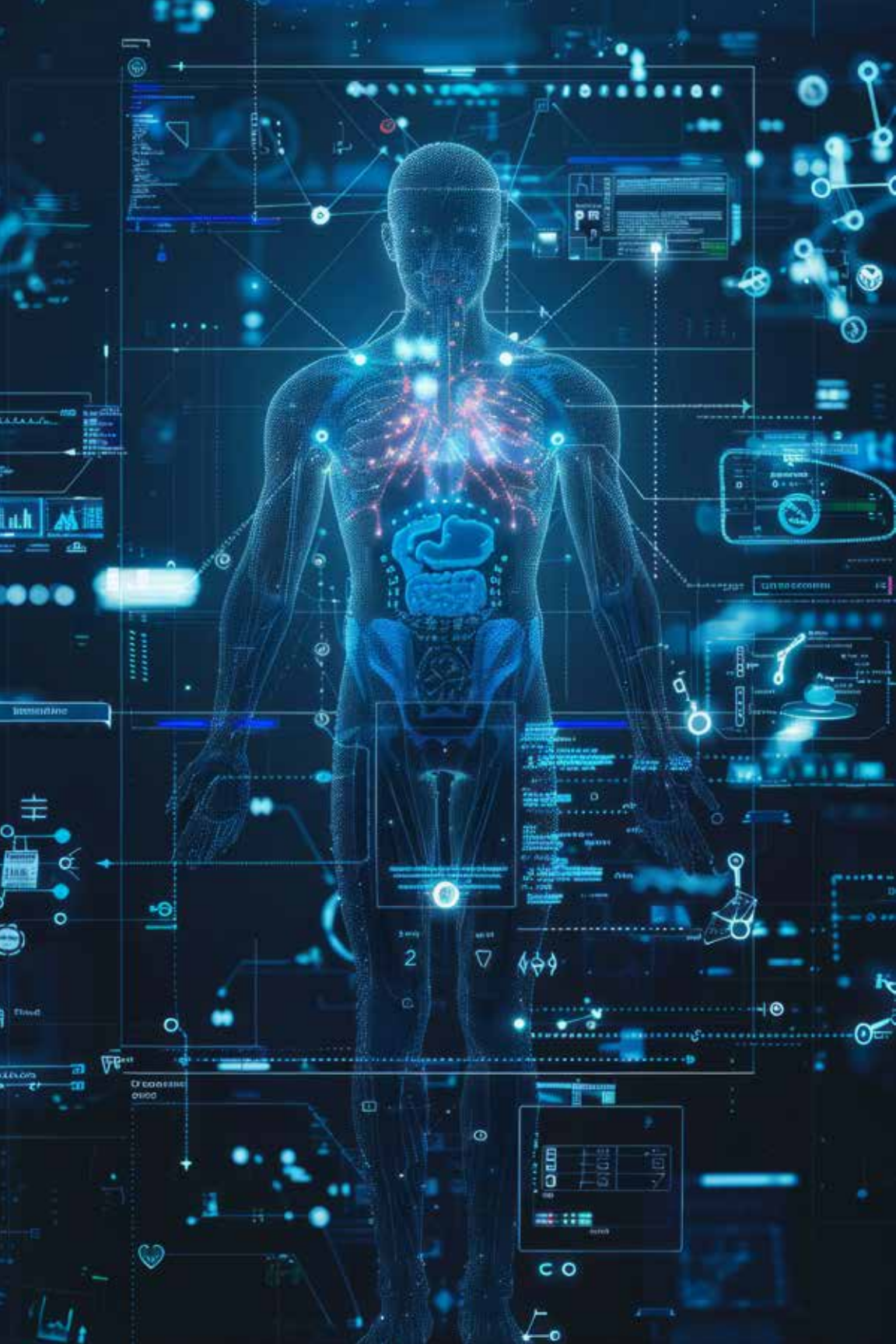
Medical professionalism; augmented intelligence; professional autonomy; automation bias; algorithmic explainability and transparency; traceability and reversibility; accountability and responsibility; physician–patient relationship; empathy and humanism.

Ideas fuerza:

- La IA debe aumentar la medicina, no sustituir al médico: el juicio clínico y la responsabilidad final siguen siendo humanos.
- El profesionalismo se juega en la autonomía y la rendición de cuentas: si el algoritmo manda o es opaco, se degrada la esencia deliberativa de la práctica.
- El mayor riesgo no es el error puntual, sino el error reproducible a gran escala y el sesgo: exige validación rigurosa y vigilancia continua.
- La relación médico-paciente puede salir ganando si la IA quita burocracia, pero se rompe si desplaza la empatía: transparencia y comunicación clara son obligatorias.
- El médico del futuro es supervisor, líder y guardián ético: debe gobernar la tecnología, resistir presiones y formarse en alfabetización digital sin perder el humanismo.

Key messages:

- AI should augment medicine, not replace the physician: clinical judgment and ultimate responsibility remain human.
- Medical professionalism hinges on autonomy and accountability: if the algorithm dictates decisions or remains opaque, the deliberative core of practice is undermined.
- The main risk is not isolated error, but large-scale reproducible error and bias: this demands rigorous validation and continuous monitoring.
- The physician–patient relationship can benefit if AI reduces bureaucracy, but it breaks down if empathy is displaced: transparency and clear communication are mandatory.
- The physician of the future is a supervisor, leader, and ethical guardian: they must govern the technology, withstand pressures, and develop digital literacy without losing humanism.



Sumario:

5 Profesionalismo médico en el contexto de la Inteligencia Artificial

1. Introducción. Profesionalismo médico ante la disrupción digital

La profesión médica históricamente se ha fundamentado en un pacto ético con la sociedad (1). Los médicos ofrecen competencia técnica, preocupación por su paciente y juicio prudente, a cambio de autonomía para ejercer en beneficio del interés del paciente (2). Este profesionalismo tradicional se ha expresado en códigos de deontología que consagran deberes de integridad, independencia y primacía del bienestar del paciente. Sin embargo, la acelerada digitalización de la salud en las últimas décadas (historias clínicas electrónicas, telemedicina, big data) ha comenzado a tensionar esos principios y prácticas (16,17).

En este contexto surge la inteligencia artificial, con un potencial transformador sin precedentes (3). La IA médica puede analizar volúmenes masivos de datos en segundos, identificar patrones ocultos y proporcionar recomendaciones diagnósticas o terapéuticas con una precisión creciente, a veces superando el rendimiento humano. Esta promesa tecnológica contrasta con preocupaciones profundas (9,12). ¿Cómo mantener la humanidad de la medicina cuando intervienen algoritmos en la toma de decisiones? ¿Cómo asegurar que el juicio clínico y la relación médico-paciente sigan siendo centrales? ¿Qué nuevos dilemas éticos y responsabilidades emergen? Estas preguntas marcan la necesidad de revisar y actualizar el concepto de profesionalismo médico en la era algorítmica.

Estos interrogantes han sido objeto de un amplio debate en la profesión médica española, cristalizando en la Declaración de Jaén sobre la Medicina 5.0 (30), documento consensuado por el Consejo General de Colegios de Médicos de España durante el X Congreso Nacional de Deontología y Ética Médica. Como miembro de esta Comisión Científica que elaboró dicha declaración, he participado en la formulación de un marco ético para integrar tecnologías emergentes preservando los principios deontológicos y el humanismo médico. Esta experiencia en la construcción de consensos profesionales sobre ética digital fundamenta el análisis crítico que desarrolla este capítulo.

2. Evolución histórica del profesionalismo médico. De la tradición hipocrática a la era digital

El concepto de profesionalismo médico hunde sus raíces en la ética hipocrática y se ha desarrollado durante siglos alrededor de determinados valores tales como el deber de anteponer el bienestar del paciente, la competencia técnica, la autonomía, la confidencialidad y la integridad moral del médico (1,2). En la era pre-digital, el médico era la fuente principal de conocimiento sanitario y gozaba de un alto grado de autoridad y autonomía en sus decisiones clínicas. Los códigos deontológicos reflejan esta tradición, por ejemplo, el Código Internacional de Ética Médica de la AMM (21) o el Código Deontológico español enfatizan que "el médico tiene el deber y derecho de ejercer con autonomía profesional e independencia clínica", siempre sirviendo a la dignidad humana y a la salud del paciente por encima de presiones externas (28). La relación médico-paciente clásica se basaba en la confianza personal, el paciente confiaba en la pericia y buen juicio del facultativo, y éste asumía plena responsabilidad por sus decisiones, combinando ciencia y prudencia clínica.

A finales del siglo XX y principios del XXI, incluso antes de la IA, el profesionalismo se enfrentó a nuevos escenarios con la llegada de la medicina basada en la evidencia, el auge de pacientes más informados y la incorporación de tecnologías informáticas (12). Se habló de nuevo profesionalismo orientado a equilibrar la autonomía del médico con la autonomía del paciente (toma de decisiones compartida), y a integrar mayores exigencias de rendición de cuentas y transparencia. La aparición de redes sociales y entornos digitales creó también el concepto de e-profesionalismo, es decir, la conducta profesional del médico en el mundo online, cuidando la confidencialidad en entornos virtuales y manteniendo la imagen profesional en redes. Todos estos cambios prefiguraron desafíos que la IA amplificaría, señalando la necesidad de que el médico del siglo XXI combine las virtudes clásicas como la excelencia, la empatía y la integridad con nuevas competencias y principios actualizados a la era digital.

En la última década se ha acelerado la digitalización de la medicina, han ido surgiendo avances que han llevado aparejados cambios en el profesionalismo. El desarrollo de la telemedicina, por ejemplo, obligó a codificar estándares para ejercer a distancia manteniendo calidad y ética (19). El Código Deontológico español de 2022 dedica un capítulo a Telemedicina,

estableciendo que es deontológicamente aceptable si se garantiza la identificación de las partes, la confidencialidad y la seguridad de la comunicación (28). La medicina de datos ha puesto en primer plano cuestiones de privacidad y uso secundario de información sanitaria, requiriendo por parte del médico conocimiento de protección de datos e implicaciones éticas del big data.

En este continuum, la IA se presenta como culminación de la transformación digital con un gran salto cualitativo (3,13). Nunca una tecnología prometió influir tan directamente en el núcleo cognitivo de la práctica médica abarcando áreas como el diagnóstico, el razonamiento clínico y el pronóstico. Por ello, múltiples instituciones han comenzado a reflexionar sobre cómo debe evolucionar el profesionalismo médico. La Asociación Médica Mundial (AMM/WMA) emitió declaraciones en 2019 y 2022 subrayando que la IA debe concebirse bajo el paradigma de "inteligencia aumentada" (21), es decir, como complemento del juicio humano y no sustituto, insistiendo en que se preserve el liderazgo del médico en la toma de decisiones y la primacía de la relación clínica.

En 2025, la AMM adoptó una declaración formal que afirma que la integración de IA "no disminuye la responsabilidad del médico por el bienestar del paciente ni su obligación de abogar por él", enfatizando que la autonomía profesional y la empatía deben mantenerse incluso con herramientas avanzadas.

En síntesis, el profesionalismo médico ha pasado de un modelo clásico centrado en la autoridad científica y moral del médico, a uno más colaborativo y tecnificado donde el médico es parte de un ecosistema de información. La era de la IA representa la próxima frontera de esta evolución. Plantea redefinir roles y competencias sin perder de vista la misión fundamental.

3. La inteligencia artificial en medicina

En la práctica médica actual, la IA se manifiesta en sistemas muy diversos: algoritmos de aprendizaje automático integrados en equipos de diagnóstico, asistentes virtuales que procesan lenguaje para resumir historias clínicas, modelos predictivos de riesgo en entornos de salud pública, e incluso chatbots capaces de ofrecer información médica.

Desde una perspectiva funcional, todas estas implementaciones pueden ser vistas como potentes herramientas diseñadas para apoyar al profesional en tareas cognitivas o administrativas. Numerosos estudios demuestran que la IA puede mejorar la precisión diagnóstica en campos como la radiología, dermatología, cardiología y anatomía patológica (4,5,14,15). Por ejemplo, se han desarrollado algoritmos de visión artificial capaces de interpretar radiografías de tórax con precisión equiparable o superior a radiólogos en la detección de ciertas patologías. Asimismo, la IA puede monitorizar constantes de pacientes en UCI en tiempo real para alertar tempranamente de deterioros, o cribar miles de artículos científicos para mantener actualizado al médico en su especialidad. Estos modelos pueden reducir errores médicos, detectar antes enfermedades, personalizar tratamientos y aliviar al facultativo de cargas rutinarias, permitiéndole dedicar más tiempo a la atención directa del paciente (3).

El carácter de herramienta de la IA implica que su valor depende del uso que se le dé. Al igual que un estetoscopio o un ecógrafo, la IA puede potenciar la capacidad del médico cuando es utilizada correcta y éticamente. Los sistemas de apoyo a la decisión clínica pueden aportar recordatorios o diagnósticos diferenciales que el médico, bajo presión de tiempo, podría pasar por alto, funcionando como una “segunda opinión” instantánea. Esto podría reducir omisiones debidas al cansancio o factores humanos complementando así la labor humana. Además, la IA puede democratizar cierto expertise (3). En entornos con escasez de especialistas, un algoritmo bien entrenado puede guiar a médicos generalistas y mejorar el acceso a diagnóstico en poblaciones remotas o vulnerables.

4. IA “copiloto” vs IA autónoma

El término copiloto se ha puesto de moda, y no sin razón. Este funciona como asistente al médico en la toma de la decisión clínica, aportando datos y sugerencias, pero dejando los controles en manos humanas. En la práctica, esto se traduce en un clínico que revisa las sugerencias aportadas, concepto physician-in-the-loop. Ninguna recomendación automatizada debería aplicarse sin la revisión y aprobación de un médico responsable (6,26). Varios marcos regulatorios apuntan en la misma dirección (7). El Reglamento Europeo de IA (AI Act) clasifica a los sistemas de diagnóstico y tratamiento como de alto riesgo, exigiendo que cuenten con características de supervisión humana, transparencia en sus recomendaciones y capacidad de explicar su lógica. El problema surge cuando las instituciones, los gestores o incluso algunos colegas empiezan a tratarlo como piloto automático.

La noción de una IA autónoma, tomando decisiones sin intervención humana, en medicina suscita inquietud ética. Si bien para tareas muy acotadas la IA puede operar sola con alta fiabilidad, delegar integralmente en la IA decisiones clínicas complejas sería equivalente a otorgarle una autoridad que hoy se asocia a la responsabilidad profesional del médico.

Un informe del Consejo de Europa (Comité de Bioética, 2021) nos advierte que una dependencia excesiva en IA puede conducir a pérdida de habilidades clínicas y a una relación médica despersonalizada (10,11). Podría surgir un nuevo paternalismo algorítmico que sustituyera al clásico paternalismo médico. Además, podría imponer recomendaciones dadas como infalibles que ni el médico ni el paciente se atrevan a cuestionar.

Lo ideal sería la colaboración entre el hombre y máquina que permita aprovechar la velocidad y consistencia de la IA sin renunciar a la responsabilidad humana.

5. Riesgos sistémicos asociados a la IA en salud

Más allá del nivel de la interacción individual médico-IA, conviene reconocer que la adopción amplia de estas tecnologías puede conllevar riesgos a nivel de sistemas de salud. Existe un riesgo sistémico que aparece cuando muchos clínicos o instituciones dependen simultáneamente de algoritmos que pudieran ser falibles. Uno de estos riesgos es el del error a gran escala (9). Si un algoritmo comercial implantado en cientos de hospitales presenta un sesgo no detectado, podría generar decisiones erróneas en miles de pacientes antes de ser corregido (8).

Históricamente, los errores médicos humanos suelen ser aleatorios y variados, pero un error algorítmico puede ser consistente y reproducible, afectando a todos los casos similares. Este potencial de daño sistemático nos obliga a redoblar los esfuerzos de validación rigurosa y vigilancia post-implementación.

Otro riesgo sistémico es la amplificación de sesgos y desigualdades en salud. La IA aprende de datos históricos, que pueden estar marcados por disparidades existentes. Podría haber menos datos de ciertas minorías y generar por tanto respuestas con sesgos. Problemas como este perpetuarían o incluso agravarían la inequidad en la atención si no se corrige. También, el acceso a la tecnología podría consolidar una medicina a dos velocidades, quienes tienen acceso obtienen beneficios, mientras poblaciones rezagadas quedan aún más atrás. Desde el profesionalismo, esto obliga al médico a abogar por una IA justa y representativa, involucrándose en la discusión sobre la calidad de los datos y sobre el acceso a los avances tecnológicos.

Es necesario también tener en cuenta que la dependencia tecnológica generalizada puede minar la resiliencia del sistema sanitario. Como hemos comentado previamente algunos profesionales pueden llegar a depender tanto de herramientas de IA que pierden destrezas básicas como el diagnóstico clínico sin ayudas. Entonces el sistema se vuelve vulnerable ante fallos técnicos. La literatura menciona este fenómeno como un empobrecimiento profesional por pérdida de habilidades (10, 11). Podríamos incurrir en el riesgo de que generaciones futuras sepan interpretar lo que sugiere el algoritmo, pero no desarrollen igual su capacidad de diagnóstico diferencial independiente. Una forma de mitigar esto, podría ser que la IA se introduzca de forma que entrene, la mente del médico y no la atrofie.

También el exceso de confianza que nos generan estas tecnologías puede ser peligroso. En centros con fuerte implantación de IA, los profesionales podrían dar por válida una decisión porque "el algoritmo lo recomienda", generando una especie de conformismo grupal. Tradicionalmente la variabilidad clínica ofrecía una cierta diversidad de opiniones que podía ser un seguro ante errores, en un entorno unificado por IA las opiniones tienden a converger en la de la máquina, reduciendo la probabilidad de detectar un fallo. Este fenómeno cognitivo es conocido como sesgo de automatización (10,11).

El profesionalismo médico, entendido como compromiso con la excelencia y la seguridad del paciente, requerirá que los médicos mantengan una actitud crítica frente a las sugerencias de IA. Los clínicos no deben renunciar a su escepticismo ilustrado y deben seguir verificando si las recomendaciones tienen sentido para el caso concreto. Si un algoritmo, sugiere un tratamiento inusual, el médico debe investigarlo, contrastarlo con la condición del paciente y, solo entonces, adoptarlo o rechazarlo.

6. Impacto de la IA en los pilares de la práctica profesional. Autonomía profesional

La autonomía profesional se refiere al derecho y la capacidad del médico de actuar según su juicio clínico y valores éticos, sin interferencias indebidas, en beneficio del paciente. Ha sido considerada un pilar del profesionalismo. Asegura que las decisiones médicas se tomen primariamente con criterios clínicos, no dictadas por intereses comerciales, administrativos o de otro tipo.

La IA puede reforzar esta autonomía al proveer al médico de información más completa y precisa para decidir (3,13). Un médico bien apoyado por herramientas inteligentes probablemente se sienta más seguro en sus decisiones y menos subordinado a guías rígidas o a la opinión de expertos remotos. La IA sería un amplificador de la capacidad del médico para tomar decisiones informadas y personalizadas. Si la IA se encarga de tareas administrativas tediosas, libera al médico de cargas poco valiosas y le facilita más tiempo para pensar y dialogar con el paciente pudiendo reducir el burnout. El clínico crece así en autonomía para dedicarse a lo que realmente aporta valor.

Sin embargo, existen preocupaciones fundadas de que la IA amenace la autonomía profesional si se convierte en una especie de autoridad técnica a la que el médico deba someterse (12). Un riesgo es la "algoritmización" de la toma de decisiones clínicas. Además podría suceder que hospitales y/o aseguradoras impongan el uso de sistemas de soporte de decisión basados en IA en aras exclusivamente de la eficiencia o uniformidad. El rol del médico podría degradarse a un ejecutor pasivo, minando la esencia deliberativa de la práctica médica (9,24).

Ya hay ejemplos incipientes. Ciertos sistemas de apoyo deciden alertas o sugerencias en la historia electrónica que los médicos se ven presionados a aceptar por política institucional. Esto ciertamente colisiona con la autonomía del clínico de decidir un plan distinto basado en su conocimiento del paciente. El Código Deontológico español, consciente de este peligro, ha consagrado que "los datos de salud extraídos de grandes bases de datos o los sistemas robóticos no sustituyen la obligación del médico de utilizar los métodos necesarios para la buena práctica" (28). La AMM igualmente reitera que el médico debe conservar la autoridad final en diagnóstico, indicación y terapia, porque su responsabilidad hacia el paciente así lo exige (21).

La clave, lógicamente, estará en mantener el equilibrio. La IA puede informar y enriquecer el juicio clínico, pero no debe imponerse de tal modo que reste la autonomía formal.

Esto nos hace replantear el concepto de decisiones compartidas. Hasta ahora ese modelo se aplicaba a la relación médico-paciente, pero es probable que haya que incluir en la ecuación también al algoritmo. Si la IA dificulta la participación del médico o del paciente, entonces está lesionando la autonomía de uno u otro.

7. Juicio médico y toma de decisiones clínicas en la era algorítmica

El juicio clínico surge de la combinación de ciencia, experiencia, intuición y conocimiento del paciente para llegar a decisiones diagnósticas o terapéuticas adecuadas. Lógicamente la introducción de IA en la ecuación hace que pueda aparecer cambios en este recorrido.

Uno de los problemas que plantea la integración de estos sistemas es que funcionan como una "caja negra". Detectan patrones y sugieren conclusiones sin proporcionar una justificación explícita comprensible para el humano. Esto contrasta con el razonamiento clínico tradicional, donde el médico suele articular las razones por las que cree en un diagnóstico como síntomas claves, hallazgos exploratorios, mecanismos fisiopatológicos. Con sistemas de IA complejos, el médico puede recibir una respuesta del tipo probabilidad de enfermedad sin más explicación.

Esta falta de explicabilidad supone un desafío doble para el juicio clínico porque el médico debe decidir cuánto confiar en esa predicción opaca (6, 25) y porque si la sigue, tal vez no pueda justificar después su decisión más que diciendo "así lo indicó el sistema". Colisiona así con la noción de que cada acto médico debe ser fundamentado. De hecho, desde una perspectiva ética, ofrecer una propuesta sin explicar cómo se llegó a ella no es correcto cuando una decisión afecta la vida o salud de una persona (28). La opacidad algorítmica amenaza con convertir el proceso deliberativo en un acto de fe en la tecnología. Esto no es propio de un profesional reflexivo.

Es razonable pensar que cuando los médicos reciben una recomendación de IA, su umbral de duda puede cambiar (36). Algunos tienden a aceptarla sin cuestionar, y pueden caer en un sesgo de automatización, mientras que otros pueden reaccionar con desconfianza excesiva con el consiguiente sesgo de aversión a algoritmos incluso siendo acertada. Ambos extremos son problemáticos. Lo ideal sería que el médico incorpore la recomendación de IA como una pieza más de evidencia y la pondere junto con la clínica.

Esto exige que el médico mantenga su razonamiento independiente activo. Un riesgo señalado es el desvanecimiento de la intuición clínica con el uso constante de IA. La intuición médica, lejos de ser algo esotérico se basa en reconocimiento de patrones sutiles fruto de la experiencia. Si

en todo momento la IA da la primera impresión diagnóstica, el médico podría perder la oportunidad de ejercitar esa "corazonada experta". Como hemos referido previamente sobre confiar en ayudas automatizadas puede llevar a la atrofia de habilidades cognitivas no utilizadas. También en el campo de la educación médica se discute cómo incorporar IA sin suprimir el aprendizaje del razonamiento clínico. Quizás usando escenarios en que el estudiante debe auditar las recomendaciones de una IA, desarrollando así pensamiento crítico.

Nuestra experiencia formativa con IA en broncoscopia (33) sugiere que estas herramientas pueden servir como complemento educativo, especialmente para profesionales en formación. Al enfrentar a residentes con las recomendaciones del algoritmo y compararlas con las decisiones de especialistas experimentados,

se genera un espacio de reflexión sobre el razonamiento diagnóstico. Los residentes que mostraron mayor concordancia con la IA (82,6%) podrían beneficiarse de analizar por qué los especialistas discrepan en ciertos casos (concordancia 65,3%), identificando así factores contextuales que las guías no contemplan. Esta metodología pedagógica, usar la IA como herramienta de confrontación y aprendizaje, podría integrarse en programas de formación especializada, enseñando simultáneamente el uso de tecnología y el desarrollo del juicio clínico crítico.

Otro aspecto para destacar es que la IA podría enriquecer el juicio clínico mediante percepciones no evidentes. Los algoritmos que integran cientos de variables podrían descubrir factores pronósticos nuevos que el médico no había considerado. En un caso real, una IA detectó que ciertos patrones en la retina predicen riesgo cardiovascular; ahora el cardiólogo informado por esa IA tiene un criterio más para juzgar el riesgo del paciente (34). Así, el conocimiento médico evoluciona con la IA, y el profesional debe actualizar su modelo mental. Vemos así el surgimiento de un juicio clínico híbrido; parte humano, parte informado por la máquina. De hecho, estudios piloto en radiología han encontrado que el dúo radiólogo + IA supera tanto a la IA sola como al humano solo en ciertas tareas de detección de cáncer. El radiólogo aporta contexto, la IA agudeza matemática, y juntos logran más aciertos (35).

Para que este modelo funcione, se necesita confianza calibrada (23). Si la IA comete errores groseros y pierde la confianza del médico, éste la descartará,

aunque podría ser útil en otras ocasiones. Por ello, la calidad y validación de los sistemas de IA es crucial para no dañar este delicado equilibrio.

Debemos aprender que introducir IA sin aval científico robusto puede arruinar su adopción e incluso empeorar la toma de decisiones por confusión. También hay implicaciones para la toma de decisiones compartida con el paciente: tradicionalmente, el médico traduce al paciente las opciones y su racionalidad. ¿Cómo explicar una decisión apoyada en IA? Los médicos deberán saber comunicar al paciente el rol de la IA.

Esto forma parte del nuevo profesionalismo. Transparencia con el paciente sobre el uso de IA, para mantener la confianza. No es aceptable esconder que una máquina decidió.

La IA está transformando el arte del diagnóstico y la decisión. El médico debe seguir tomando decisiones activamente, ahora con una herramienta más potente. Debe cultivar una actitud de colaboración crítica, ni servilismo al algoritmo ni rechazo apriorístico, sino interrogar a la IA, contrastar con su conocimiento y con la singularidad del paciente.

5 Profesionalismo médico en el contexto de la Inteligencia Artificial

8. Impacto de la IA en la relación médico-paciente

La relación médico-paciente es frecuentemente descrita como el corazón de la medicina. Es un vínculo humano basado en la confianza, la comunicación sincera, la empatía y el respeto a la autonomía del paciente. A lo largo de la historia, cada avance tecnológico ha suscitado preguntas sobre cómo afecta en ese encuentro clínico interpersonal. La IA no es la excepción; de hecho, por su capacidad de intervenir en decisiones y procesar información del paciente, la transformará de una manera profunda.

Paradójicamente, una herramienta muy avanzada podría hacer la atención más humana si se usa correctamente. Si una IA automatiza la redacción del historial o extrae datos relevantes del registro, el médico ya no tendrá que teclear durante la consulta y puede dedicar más atención visual y mental al paciente. Modelos de software que actúan como escribas están siendo implantados en muchos centros sanitarios. Actualmente los médicos de atención primaria pasan hasta el 50% del tiempo mirando la pantalla en vez de al paciente (15). Un médico liberado de tareas administrativas tiene más espacio para escuchar activamente, que es justamente lo que los pacientes más valoran en una consulta.

Otra ventaja potencial es la accesibilidad y continuidad en la relación. Asistentes virtuales podrían responder dudas de la salud del paciente fuera de consulta. Si estos sistemas están bien integrados, el paciente siente un acompañamiento más continuo, y el médico puede intervenir cuando detectan una alerta. Bien gestionado, esto refuerza la alianza terapéutica, porque el paciente percibe que su equipo de salud, humano y digital está pendiente de él. En enfermos crónicos, la IA puede monitorear su adherencia o parámetros clínicos y notificar al médico de la necesidad de contacto. El paciente recibe atención en el momento oportuno, fortaleciendo la sensación de cuidado.

Además, la IA puede ayudar a educar al paciente traduciendo jerga médica a lenguaje sencillo, mostrando visualizaciones de su estado de salud, etc. En teoría, esto facilita una comunicación más clara y un paciente más involucrado. Incluso hay desarrollos incipientes de IA que detectan emociones en la voz o gestos del paciente. Algún día podrían alertar al médico si perciben que el paciente está confundido, escéptico o triste, de manera que el clínico atienda esa dimensión emocional que quizá estaba pasando por alto. Por

tanto, bajo una óptica optimista, la IA puede ser un facilitador de la relación médico-paciente.

Existen también preocupaciones respecto a los planes de implantación de esta tecnología. Una de las principales es la deshumanización (27). Ya se notó en la era digital previa, desde entonces los pacientes se quejan de que miran más a la pantalla que a sus ojos. Con la IA este riesgo puede continuar, si el médico confía exclusivamente en métricas y predicciones, puede prestar menos atención a la narración personal del paciente. Se perdería la escucha activa y la validación del paciente como persona. Eliminar ese espacio relacional que constituye la esencia de la medicina es un error irreversible.

La relación clínica es un documento emocional que se escribe con las dudas, miedos y esperanzas del paciente en interacción con el profesional. La IA no sabe leer ni respetar ese "texto"; solo el ser humano puede. Por tanto, si la IA absorbe la atención del médico, la relación se empobrece.

Otro problema podría ser la erosión de la confianza. La confianza del paciente en su médico se sustenta en creer que este está actuando con competencia, benevolencia y de forma individualizada. Si los pacientes perciben que las decisiones vienen dictadas por una máquina, podrían dudar de si el médico realmente está ejerciendo juicio o simplemente siguiendo un protocolo impersonal. Un estudio en JAMA (29) encontró que los pacientes aceptaban textos escritos por IA, pero cuando se les informaba que había IA detrás, su satisfacción disminuía ligeramente. Esto sugiere que la transparencia sobre la IA es un arma de doble filo, es ética y necesaria, pero puede influir en la percepción.

Para manejarlo, los médicos deberán educar y tranquilizar a los pacientes sobre la IA explicando que es una herramienta revisada por ellos y que no substituye su criterio. La honestidad será clave: si un paciente pregunta "¿Doctor, es usted o un ordenador quien decide mi tratamiento?", el médico debe explicar el proceso claramente, enfatizando su rol garante. De hecho, la regulación europea insiste en el derecho del paciente a no ser objeto de decisiones totalmente automatizadas sin intervención humana, reforzando que legalmente el paciente puede exigir intervención médica real (7).

Un punto crítico es la empatía. Un médico descargado de tareas repetitivas podría tener mejor ánimo y más tiempo para mostrar empatía. Pero si el

clínico depende mucho de datos y pantallas, podría caer en una aproximación más fría, olvidando la dimensión emocional. La empatía requiere presencia plena y atención a las claves sutiles: tono de voz, expresiones faciales. Elementos que nada tienen que ver con algoritmos. Es improbable que ninguna IA supla eso, aunque existen desarrollos de IA con respuestas empáticas simuladas, la genuina relación humana es insustituible. La calidez, la compasión, la mano en el hombro en un momento difícil son actos profundamente humanos que construyen relación terapéutica y que nunca podrá replicar un software (28).

Por todo ello, la IA introduce nuevas dinámicas de relación. La confidencialidad y privacidad también atañen a la relación; el paciente confía datos íntimos al médico; si ahora también se los confía a un sistema de IA, ¿quién garantiza la privacidad? Por ello, un profesionalismo robusto en la era digital incluye defender la privacidad del paciente ante la IA, asegurarse de usar sistemas aprobados, con medidas de anonimización, y con consentimiento informado específico cuando aplique. Esto debe ser comunicado al paciente para que su confianza en el médico se extienda a la tecnología porque confía en el médico que la maneja.

El vínculo médico-paciente puede salir fortalecido o debilitado con la IA, dependiendo de cómo se integre. Si se hace centrada en el paciente, respetando la autonomía y la necesidad de comunicación clara, la IA puede liberar tiempo y aportar información útil, profundizando la relación. Pero si se implementa con afán de productividad o como sustituto de la interacción humana, hay riesgo de fractura en la confianza y despersonalización.

9. Nuevos roles del médico en la era de la IA

A medida que la IA se integra en la atención sanitaria, el perfil competencial y de rol del médico está experimentando cambios significativos. Más que desaparecer, la figura del médico se transforma y diversifica. A las funciones clásicas de diagnosticar, tratar y confortar, se suman ahora responsabilidades de

supervisión tecnológica, liderazgo en equipos multidisciplinares y garante de la ética en entornos automatizados.

10. **Médico supervisor de la IA**

La primera, y quizá más inmediata, nueva función es la de supervisor o controlador de la inteligencia artificial clínica (6,25). Lejos de imaginar un futuro en el que los médicos sean reemplazados por máquinas, las recomendaciones internacionales insisten en mantener siempre un médico responsable de la toma de decisiones asistida por IA. Esto convierte al profesional en un vigilante activo del desempeño de los sistemas inteligentes. Su tarea es revisar, validar o rectificar las sugerencias de la IA antes de aplicarlas al paciente.

Nuestra investigación sobre triaje avanzado en urgencias (31) ilustra este principio de supervisión activa. Al implementar ChatGPT-4.0 para clasificación de pacientes, demostramos que la precisión del algoritmo (índice Kappa de 0,81) dependía fundamentalmente de tres elementos controlados por médicos: el diseño de prompts estructurados, la selección de variables clínicas relevantes, y la validación continua del output. Sin esta supervisión humana cualificada, el mismo algoritmo producía resultados inconsistentes. Este hallazgo refuerza que la IA en medicina no funciona como herramienta autónoma, sino como extensión de la capacidad del médico que la configura, supervisa y valida.

Ser supervisor de IA implica desarrollar habilidades nuevas. Se requiere una comprensión básica del funcionamiento del algoritmo para poder interpretar sus salidas, y tener la capacidad de detectar cuándo algo no cuadra entre la recomendación de la IA y la realidad clínica, lo que requiere experiencia y juicio. Un buen supervisor humano conoce las fortalezas y debilidades de su sistema. Es deber del médico conservar el juicio profesional y la responsabilidad final sobre cada diagnóstico e indicación. Esto equivale a situar al médico como garantía de calidad, su revisión es la última barrera para evitar que un error algorítmico alcance al paciente.

Se vislumbra la aparición de especialistas en seguridad algorítmica dentro de los hospitales, médicos con formación adicional en análisis de datos, que ayuden a monitorear los algoritmos en funcionamiento. Se recomiendan auditorías periódicas de los algoritmos en la práctica para comprobar que siguen funcionando con la precisión esperada, y los médicos serán parte de esas auditorías, aportando retroalimentación basada en casos reales.

Este rol de supervisor conecta con el concepto de medicina basada en la evidencia, ahora potenciada por la IA. El médico verifica que la recomendación algorítmica esté alineada con la evidencia clínica y con la situación individual. Si encuentra discordancia, debe tener la autoridad y la confianza institucional para anular la sugerencia automática. Como antes un médico podía cuestionar una guía o indicación superior por el bien del paciente, ahora debe sentirse empoderado para cuestionar a la IA cuando sea prudente. Idealmente, la organización sanitaria debe respaldar esta función, evitando cualquier cultura de "el ordenador tiene siempre razón".

Este nuevo papel también significa responsabilidad proactiva (23). El médico no espera a que la IA falle para intervenir, sino que anticipa y previene errores. Si un modelo diagnóstico no es entrenado en ciertas poblaciones, un médico supervisor podría advertir de "no utilizar en pacientes pediátricos, riesgo de fallo", o "necesitamos entrenar un nuevo modelo para este grupo". Este tipo de input del médico a los desarrolladores es crucial. En lugar de un usuario pasivo, el médico se convierte en una codesarrollador in situ, calibrando el sistema con su feedback. Muchos ingenieros de IA en salud reconocen que sin la retroalimentación continua de los clínicos, sus modelos no pueden mejorar ni adaptarse a la complejidad real. El médico supervisor de IA es la evolución del médico como garante de la calidad asistencial. Antes vigilaba el desempeño de sus residentes o de su equipo, ahora ha de vigilar también a sus herramientas inteligentes. Es un cambio de escenario, pero en el fondo es una extensión del rol ético de non-maleficencia.

11. Médico líder de sistemas complejos

Otra faceta emergente es la del médico como líder y gestor en la implementación de sistemas que involucran IA. La prestación sanitaria con IA deja de ser un acto individual para ser un proceso de equipo multidisciplinar: médicos, ingenieros de datos, informáticos, gestores y a veces la propia voz del paciente. En este escenario, el liderazgo clínico es indispensable para asegurar que la tecnología se traduzca en mejoras reales en salud. Se ha visto en múltiples intentos de digitalización que, sin la apropiación por parte de los clínicos, las innovaciones fracasan o generan rechazo. El médico líder de sistemas complejos actúa como puente entre la tecnología y la práctica clínica (3,12). Entiende las necesidades del frente asistencial

y puede comunicarlas al equipo técnico, y viceversa, traduce las posibilidades y limitaciones de la IA a sus colegas sanitarios. Este rol de liderazgo implica también una responsabilidad pedagógica. Actúa como agente de cambio, contrarrestando resistencias comprensibles al cambio tecnológico mediante la demostración de beneficios y la escucha de las preocupaciones de los compañeros.

Asimismo, el liderazgo clínico se extiende a aspectos de gobernanza. El médico debe participar y encabezar comités éticos, comités de selección de tecnología y elaboración de protocolos relacionados con IA. Es fundamental que la voz clínica esté presente al decidir qué sistema de IA comprar, cómo validarlo, cómo monitorizar su rendimiento y qué hacer si produce un resultado conflictivo. Los organismos colegiales o sociedades científicas por especialidades deberán elaborar códigos de práctica para la IA y actualizar constantemente las líneas éticas y profesionales a medida que la tecnología evoluciona. Esto sugiere que los médicos líderes participarán no solo a nivel hospitalario, sino también en organismos profesionales redactando recomendaciones de alcance nacional e internacional. Ser líder en un sistema complejo con IA también entraña coordinar un equipo interprofesional. Este debe ser un liderazgo muy colaborativo y transversal, diferente del clásico liderazgo jerárquico médico dentro de su silo. Requiere habilidades de comunicación con profesionales no médicos, entendimiento de conceptos de gestión de proyectos, e incluso algo de conocimiento legal/ regulatorio. En cierto modo, es un rol de médico gestor de innovación.

Una dimensión no menor es que el médico líder debe asegurar que la voz del paciente también se considere en la implementación de IA. Puede implicar consultar a asociaciones de pacientes sobre la aceptabilidad de ciertas herramientas, o incluir medidas de experiencia del paciente en la evaluación del proyecto. Esto es parte del liderazgo en el sentido de mantener la brújula centrada en la misión.

El médico líder de sistemas con IA es como un director de orquesta que mantiene la armonía entre tecnología, equipo humano y objetivos asistenciales. Para formar estos líderes, serán necesarios programas específicos y es deseable que estos médicos tengan reconocimiento institucional, ya que su trabajo muchas veces es invisible. Este liderazgo médico es crucial para evitar una posible tecnocracia sanitaria asegurando que siempre haya liderazgo clínico guiando la IA con criterios científicos y éticos propios de la profesión médica.

12. **Médico garante ético y humanista**

El tercer rol emergente es, quizá, el más intangible pero fundamental. El médico como guardián de la ética y los valores humanistas en un entorno crecientemente dominado por datos, automatización y lógica algorítmica. Tradicionalmente, los médicos ya son depositarios de un código deontológico y se espera de ellos un comportamiento ético individual. Pero ahora se requiere algo más, una vigilancia ética proactiva sobre cómo se emplean las nuevas herramientas para proteger principios como la dignidad humana, la equidad, la compasión y la privacidad (19,21).

La IA, por sus características, puede tender a despersonalizar la atención si no se pone consciencia ética. Los algoritmos operan sobre poblaciones y probabilidades, mientras que la ética médica exige ver al individuo irrepetible con sus derechos y valores. El médico debe ser el contrapeso moral que recuerde siempre que detrás de cada dato hay una persona, y que ninguna optimización algorítmica justifica violar la dignidad o autonomía de un paciente. Una IA podría predecir con alta probabilidad que cierto paciente no se beneficiará de un tratamiento costoso, y un gestor podría decir que, según la IA, no le ofrezcan terapia. El médico como garante ético intervendrá para asegurar que las decisiones clínicas se individualicen, y que incluso si la probabilidad es baja, se consulta con el paciente su preferencia y se respeta su derecho a un trato no discriminatorio. Como subraya el Código Deontológico español, los avances de IA y big data deben realizarse en beneficio de la sociedad y la salud pública, lo cual incluye la responsabilidad de no exacerbar desigualdades ni comprometer valores fundamentales (28).

Otro aspecto es la transparencia y honestidad. El médico insistirá en que los pacientes tengan derecho a saber cuándo su atención involucra IA y a recibir explicaciones comprensibles. Si surgen riesgos de privacidad, él debe alzar la voz para corregir el rumbo. Este rol de guardián también se refleja en la denuncia de sesgos. Si un médico observa que el algoritmo trata sistemáticamente peor a un grupo, tiene un deber ético de señalarlo a la institución o desarrollador para subsanarlo. En cierto modo, los médicos se convierten en observadores críticos del comportamiento de la IA y defensores de los pacientes ante posibles injusticias producidas por automatización. Esto conecta con la idea de justicia dentro del profesionalismo. El médico no solo atiende a quien tiene delante, sino que tiene un compromiso con la sociedad y con la equidad en el sistema sanitario.

Además, cabe mencionar la empatía y el humanismo. El médico se ocupa de que, a pesar de toda la tecnología, el trato humano no se degrade (12). En la era algorítmica, donde es fácil obnubilarse con eficiencias y números, esta perspectiva es vital para no perder el norte humanístico de la medicina. A veces, ser garante ético significará abogar por excepciones a la regla técnica en nombre de la compasión. Ese acto humanitario es parte del profesionalismo, y ningún algoritmo puede captarlo. La figura del médico como garante ético también juega un papel hacia sus colegas, actuando como referente moral en su equipo, fomentando la reflexión bioética sobre la IA. Mantiene viva la deliberación ética en torno a la IA para que no se normalicen prácticas cuestionables por mera costumbre o presión.

13. Profesionalismo ante nuevos desafíos

Habiendo explorado cómo cambian los roles y relaciones del médico con la IA, abordamos ahora tres desafíos transversales que tensan los valores profesionales: la opacidad de los algoritmos y su necesidad de explicabilidad, las

presiones organizativas y económicas asociadas a la adopción de IA, y la evolución de la medicina defensiva en un contexto algorítmico. Cada uno encierra riesgos de desviación del profesionalismo si no se manejan con cuidado.

14. La exigencia de transparencia y explicabilidad

Buena parte de la IA actual, en especial las redes neuronales profundas, opera como una verdadera caja negra. Reciben datos de entrada y entregan un resultado, pero el proceso interno es tan complejo que ni los propios creadores pueden explicar fácilmente cómo llegaron a esa conclusión. En medicina, aceptar ciegamente una decisión sin comprensión va contra la tradición racional y ética.

Desde los juramentos hipocráticos hasta hoy, se espera que el médico sepa por qué hace lo que hace y pueda justificarlo. Un algoritmo opaco desafía

esta capacidad de justificar sus decisiones, y por ende choca con el profesionalismo que demanda rendición de cuentas. Desde la ética médica y la regulación se clama por la explicabilidad de la IA en salud. No se espera que cada médico entienda las matemáticas profundas del modelo, pero sí que disponga de información inteligible sobre los factores que influyeron en la recomendación algorítmica. Un sistema de IA diagnóstico debería poder resaltar qué variables del paciente pesaron más para sospechar tal enfermedad. Si un sistema sugiere infarto y puede señalar que se ha basado en alteraciones en ECG y enzimas elevadas, el médico tiene algo con qué validar. Esto hace a la herramienta más útil y segura.

El Código Deontológico español (28) recogió este principio al indicar que "el médico debe exigir un control ético y finalista de la investigación con IA basado en la transparencia, reversibilidad y trazabilidad de los procesos". La transparencia y trazabilidad equivalen a explicabilidad, poder seguir la ruta de cómo la IA llegó a su output. La reversibilidad sugiere que debe ser posible corregir o deshacer acciones de la IA si se descubre un error, otro aspecto importante.

Desde el punto de vista del profesionalismo, el médico debería negarse a usar sistemas totalmente opacos para decisiones críticas sin al menos una validación externa sólida. Y si se ve obligado por falta de alternativa, debe extremar la supervisión. Las instituciones tienen el deber de proporcionar herramientas auditables. El algoritmo que no sea explicable, no debería ser aplicable (6,7,19). Esto puede chocar con intereses comerciales de empresas que no quieren revelar su receta secreta. Ahí es donde el médico como colectivo debe abogar por el derecho a la explicación en salud, incluso si se requiere presión regulatoria. El Reglamento de la UE (7) exige transparencia para sistemas de alto riesgo sanitario. Esta claridad empodera al médico para ejercer con criterio.

Si sabe por qué la IA sugiere algo, puede juzgar si es razonable o si hay un error. También le permite explicar al paciente la decisión, manteniendo la confianza. Y finalmente entrena al médico y mejora su conocimiento, una IA explicable puede mostrar correlaciones novedosas, lo que enriquece la base de conocimientos del profesional. Ahora bien, la realidad actual es que muchas IA no son fácilmente explicables. El médico no ha de negarse necesariamente a su utilización por ello. Debe hacer balance. Si un modelo es muy preciso y salva vidas, quizá se use, aunque sea un poco opaco,

pero siempre con la intención de limitar al máximo el desconocimiento del modelo.

Otro punto es la calidad de los datos. Transparencia también implica conocer con qué datos fue entrenada la IA. Un médico profesional debería preguntarse: ¿este algoritmo fue probado en pacientes como el mío? ¿tiene sesgos conocidos? La respuesta ideal vendría en la documentación técnica del producto. Si no viene, es una bandera roja. El profesionalismo aquí demanda un sano escepticismo, no asumir que toda IA es neutra, sino investigar su origen. El médico no debe comportarse como un simple usuario sometido; es un guardián de la calidad que tiene derecho a saber y entender las herramientas que utiliza. De ser necesario, puede rechazar tecnologías que comprometan la seguridad por su opacidad. La tecnología puede ser muy inteligente, pero lo verdaderamente inteligente es usarla con prudencia.

15. Presiones organizativas y económicas

En la práctica clínica, los médicos no operan en el vacío, sino dentro de estructuras sanitarias como hospitales o centros de salud y bajo marcos económicos públicos o privados. La introducción de IA ocurre en ese escenario real, y puede venir acompañada de presiones organizativas y económicas que ponen a prueba la independencia y los valores profesionales del médico.

Las organizaciones sanitarias, desde gerencias de hospital hasta sistemas nacionales de salud, ven en la IA un potencial para mejorar la eficiencia y reducir costes (24). Nuestra investigación sobre flujos de trabajo asistidos por IA en triaje de urgencias (32) cuantificó este potencial mediante simulación de dos circuitos: el tradicional (triaje→médico→prueba) versus uno alternativo (triaje→prueba→médico). La activación temprana de pruebas diagnósticas basada en recomendaciones de ChatGPT-4 redujo el tiempo medio de atención en 14,4 minutos por paciente, con mayor impacto en hospitales públicos saturados (ahorro superior a 22 minutos) y menor en contextos privados de baja demanda (6,3 minutos). Este hallazgo confirma que la IA puede generar ganancias operativas significativas, especialmente en entornos de alta presión asistencial donde los retrasos diagnósticos son un factor crítico de saturación.

Sin embargo, esto es válido y positivo siempre que la eficiencia redunde en más pacientes atendidos y mejor atención, no en reducción de calidad. La historia muestra que a veces las administraciones persiguen eficiencias económicas a costa de la calidad asistencial. Existe la preocupación de que directivos puedan imponer el uso de ciertas IA para ahorrar, sin considerar adecuadamente implicaciones éticas o de seguridad. Este tipo de presiones generaría un conflicto de lealtades entre la lealtad al paciente (principio de primacía del bienestar del paciente) y la lealtad a la institución/empleador que exige seguir la herramienta por coste-efectividad.

El profesionalismo médico resuelve este conflicto tradicionalmente poniendo al paciente primero. De hecho, los códigos de ética son claros en que el médico debe mantener su independencia clínica incluso ante presiones comerciales o administrativas, y denunciar si se le coacciona a actuar contra el interés del paciente. La IA no cambia ese deber; lo actualiza. Si el médico siente que la IA organizativa perjudica al paciente, debería poder objetar. Obviamente, esto coloca al médico en situación incómoda dentro de su institución. Se requerirá valentía profesional y respaldo colegial para sostener posturas así. Los colegios de médicos tendrán que apoyar a quienes enfrenten presiones indebidas de seguir al algoritmo por decreto cuando haya base para dudar de su adecuación.

Otro frente de presión puede ser más sutil: la cultura corporativa de la innovación. Muchos hospitales se sienten empujados a adoptar IA para no quedarse atrás. Se podría generar un clima donde cuestionar la IA se vea como resistencia al cambio mal vista. Un profesionalismo bien entendido debe distinguir entre resistencias retrógradas e interrogantes legítimos. El médico que plantea que un algoritmo no fue validado adecuadamente o que puede causar inequidades, está actuando éticamente, no siendo terco. Sin embargo, es previsible que algunos pioneros entusiastas tilden a escépticos de obstaculizar el progreso. Aquí los líderes médicos deben fomentar un ambiente donde se puedan plantear dudas sin represalias, y las decisiones de implementación se tomen con base científica y ética, no por modas o imposiciones de directivos que tal vez no comprenden las sutilezas clínicas (37).

En términos económicos, la IA también trae la cuestión de la responsabilidad financiera. Si un error de IA ocasiona gasto extra, ¿lo cubre el sistema o recae en el departamento/clínico? Esto podría crear incentivos perversos, quizá un hospital argumente que, dado que su algoritmo reducirá costos, se deben

seguir sus recomendaciones y si se desvía y se consumen más recursos, tendrá que justificarse exhaustivamente. Eso puede cohibir al médico a tomar la ruta más barata indicada por IA, incluso si clínicamente sospecha que el paciente necesita la opción más costosa. Se corre el riesgo de que la IA sirva como herramienta de gerenciamiento de costos, pasando por encima del juicio clínico personalizado. En contrapartida, si el médico sabe que el algoritmo a veces se equivoca a favor de ahorro, podría practicar medicina defensiva al revés, es decir, ignorar al algoritmo para no infra tratar y luego tener problemas si algo va mal. Ambas reacciones revelan cómo la presión económica puede distorsionar la toma de decisiones.

El profesionalismo demanda que las decisiones clínicas se tomen fundamentalmente por criterios clínicos y éticos, y solo secundariamente se consideren costos, dentro de lo razonable. Los médicos, con visión de bien común, sí deben ser conscientes de los recursos, pero nunca de forma que comprometa la calidad individual de la atención. Si la IA sugiere ahorrar, bien, pero solo si es clínicamente adecuado.

Existen implicaciones a nivel macro. Los médicos, a través de sus colegios y sociedades científicas, deben participar en cómo se incorporan las IA en las políticas de salud. Si el sistema nacional quiere usar IA por ejemplo para priorizar listas de espera quirúrgicas, habrá consideraciones éticas y clínicas a debatir. Los médicos como representantes de los pacientes tienen el deber de incidir en esas decisiones para que primen la justicia y la ética clínica sobre meros criterios económicos.

Un último factor son las presiones de mercado e industria. Las empresas de tecnología verán en los médicos prescriptores o influenciadores para adoptar sus IA. Podría aparecer un conflicto de interés. Esto es análogo a lo que ocurre con fármacos. El profesionalismo ya lidia con ello estableciendo transparencia de conflictos y primacía del interés del paciente. Se deberán extender esos mismos controles a la interacción con empresas de IA, códigos de conducta para médicos que colaboran con la industria tech, revelar vínculos financieros, etc., para que las recomendaciones se basen en evidencia y no en interés propio. También, evaluar críticamente estudios patrocinados sobre IA antes de cambiar prácticas.

Las presiones organizativas y económicas alrededor de la IA exigen que el médico reafirme su independencia clínica y su rol de abogado del paciente. No se trata de obstinación irresponsable, el médico debe colaborar con

la eficiencia del sistema, pero nunca a costa de la calidad o equidad en la atención individual. Este equilibrio no es sencillo y requerirá fortaleza moral y apoyo colegial para médicos que se vean entre la espada, la gerencia y la pared, el deber hacia su paciente. El mensaje central es que la IA debe ser una herramienta al servicio de la medicina, y no la medicina al servicio de la herramienta o de quien la patrocina. Mantener esa jerarquía de valores será una prueba definitoria para el profesionalismo médico en esta década.

16. **Medicina defensiva aumentada por IA**

La medicina defensiva se refiere al fenómeno de tomar decisiones médicas no tanto para beneficiar al paciente, sino para que el profesional se proteja legalmente. Podría pensarse que la IA, al ser más precisa y objetiva, reduciría errores y demandas, aliviando la medicina defensiva. Pero también podría generarse una nueva forma de medicina defensiva relacionada con la IA (24,25).

Una posible manifestación es la defensiva por adherencia al algoritmo: el médico sigue ciegamente lo que la IA sugiere para, en caso de resultado adverso, poder decir "yo hice lo que el sistema indicaba, no tengo culpa". En lugar de usar su criterio, se refugia en la supuesta autoridad del algoritmo. Esta conducta es peligrosa, pues sacrifica el juicio clínico y la personalización. Otra variante es la defensiva por sobreuso de IA. En este caso el clínico ordena toda herramienta algorítmica disponible, no porque la necesite para el caso, sino para poder decir que agotó todos los recursos. Esto encarece la atención innecesariamente y puede generar confusión. Es similar a pedir pruebas de más, solo que ahora con softwares.

También existe la faceta contraria, la defensiva por rechazo de IA. Si un médico teme que al usar IA puede equivocarse y ser demandado si, por ejemplo, el algoritmo falla y a él le culpan, podría decidir no usarlo, aunque esté disponible y recomendado. Prefiere fiarse solo de su criterio porque al menos así controla la situación y sabe cómo justificarlo. Si bien la prudencia es comprensible, negarse a usar una herramienta validada que mejora resultados podría considerarse negligente en el futuro.

5 Profesionalismo médico en el contexto de la Inteligencia Artificial

Existe además un aspecto psicológico. La delegación de culpa. Un médico con IA podría, consciente o inconscientemente, culpar al algoritmo en caso de error. Si bien en términos legales actuales eso no exime al médico, puede crear una cultura de menor asunción de responsabilidad moral individual. El desafío será mantener la accountability, si ocurrió un daño, ver qué parte fue fallo humano y qué parte fallo del sistema, y aprender ambos. Sin una cultura sana, se corre riesgo de traspasar la culpa; médicos culpando a ingenieros, ingenieros diciendo que los médicos no supervisaron bien. Para el paciente eso es terrible, porque siente indefensión ante un sistema en que nadie asume culpa. Por eso, a nivel macro, el profesionalismo médico también se extenderá a demandar claridad en la atribución de responsabilidades para no caer en ese vacío. Si los clínicos confían en que no serán penalizados injustamente por un error algorítmico inadvertido, estarán más dispuestos a usar IA sin comportamientos defensivos.

Es fundamental involucrar a los pacientes en la comprensión de la IA. Los pacientes deben entender que la medicina no es infalible, ni con IA. La transparencia y diálogo podrían generar más confianza mutua que desincentive la actitud defensiva del médico. Un médico que entiende por qué la IA recomienda algo puede explicar y justificar mejor su actuar, y tendrá menos miedo a que parezca una decisión irracional ante otros.

La medicina defensiva es un mal conocido que la IA podría agravar si no se encara. El profesionalismo médico deberá incluir la reflexión sobre estos comportamientos y promover entornos donde los médicos no sientan que su principal preocupación es cubrirse legalmente, sino curar con seguridad. Esto requerirá ajustes en la cultura de la práctica clínica. Si se logra, la IA podrá desplegar sus beneficios sin engendrar más miedo ni excesos de precaución contraproducentes.

17. **Alfabetización en IA y formación médica**

La alfabetización en inteligencia artificial es ya considerada una competencia esencial para los profesionales de la salud del siglo XXI (22). No significa que cada médico deba ser programador, pero sí que entienda los fundamentos de qué puede y qué no puede hacer la IA, cómo aprende, qué es un sesgo algorítmico, y los principios éticos y de seguridad asociados. En la educación médica, esto representa un nuevo paradigma formativo.

En el pregrado, varias instituciones están empezando a incluir en sus currículos contenidos sobre salud digital e IA. En España, se debate incorporar materias optativas o transversales sobre big data, informática médica y ética de la tecnología. Entender la lógica de un algoritmo no es tan distinto de entender la de una prueba diagnóstica, ambas tienen métricas de desempeño y limitaciones. Formar en IA complementa la formación en pensamiento crítico y método científico.

La Organización Mundial de la Salud (OMS) en su orientación sobre la IA en salud (2021) recomendó que los países integren capacitación en IA en los planes de estudio de ciencias de la salud, enfatizando la interdisciplinariedad y la ética. Asimismo, la UNESCO en su Recomendación de Ética de IA (2021) sugiere capacitación de los profesionales para poder interactuar con estas tecnologías de forma informada y crítica (19,20). Un contenido indispensable es la gestión de datos y la privacidad. Los médicos deben saber la importancia de la calidad de datos para entrenar IA, los peligros de datos mal recogidos, y las regulaciones de protección de datos aplicables. Esta conciencia de ciberseguridad y confidencialidad debe ser parte del profesionalismo digital.

Otra competencia a desarrollar es la interpretación crítica de literatura sobre IA. Cada vez más artículos en revistas médicas reportan resultados de algoritmos. El profesional debe poder leer esos estudios con ojo crítico, entender métricas, saber si un estudio de validación es robusto o tiene sesgos, etc. En suma, aplicar sus habilidades de lectura crítica de evidencia también a este campo (23). Solo así podrá juzgar qué herramientas valen la pena incorporar a su práctica. La alfabetización también habilita a los médicos a participar en equipos de desarrollo. Un médico con nociones puede colaborar mejor con ingenieros para cocrear soluciones pertinentes. Actualmente se observan médicos aprendiendo programación básica o análisis de datos para contribuir en investigación de IA.

No hay que olvidar la dimensión comunicativa. La formación debe incluir cómo comunicar a los pacientes sobre la IA. Esto es algo relativamente nuevo. Entrenar a médicos a explicar en qué grado se ha empleado IA y sus consecuencias. Que lo puedan hacer de forma sencilla, honesta y tranquilizadora. Es necesario un cambio cultural para que los médicos vean a la IA no como amenaza sino como otra herramienta a dominar. La mejor forma es a través del conocimiento. Históricamente, la profesión médica ha incorporado avances como la radiología o la genética adaptando formación y generando subespecialistas. Con la IA ocurrirá similar, habrá subespecialistas muy enfocados, pero todos necesitarán un grado general de competencia. Alcanzar esto es parte de la responsabilidad de las instituciones formativas y colegiales. De lo contrario, se corre el riesgo de brechas generacionales.

18. Conclusiones críticas

A la luz de todo lo analizado, podemos concluir que nos encontramos en un punto de inflexión histórico para el profesionalismo médico. La irrupción de la inteligencia artificial está transformando herramientas, procesos y roles, pero también está poniendo a prueba los valores atemporales de la profesión.

En términos prácticos, cambia la forma en que se ejerce la medicina. Los algoritmos pueden asumir muchas tareas cognitivas rutinarias, liberando al médico de ciertas cargas, pero también obligándolo a un nuevo rol de gestor de información y supervisor tecnológico. Cambian las competencias requeridas. Cambia también la dinámica de decisión, de una relación médico-paciente tradicional pasamos a un ecosistema médico-paciente-IA, lo cual añade complejidad en la comunicación y en la toma de decisiones compartidas. Se produce un cambio en la relación del médico con su propio conocimiento y autoridad.

A pesar de los enormes avances, los principios éticos y humanísticos básicos de la medicina deben permanecer inmutables. La primacía del paciente, su dignidad, sus valores, su bienestar, deben seguir estando en el centro. Tampoco cambia el deber de compasión y empatía (28). Debe también

permanecer la obligación de excelencia profesional. El médico sigue teniendo el deber de mantenerse actualizado y competente.

Escribo esto después de haber pasado muchos años viendo cómo la tecnología cambia la medicina, a veces para mejor, a veces no tanto. La IA no es diferente de otras revoluciones que hemos vivido. La digitalización de las historias clínicas, la telemedicina, los sistemas de alerta. Todas prometieron maravillas. Todas trajeron problemas no anticipados. La diferencia es que esta vez la tecnología interviene en el núcleo mismo de lo que hacemos: pensar, decidir, diagnosticar. Por eso no podemos permitirnos ser espectadores pasivos de este cambio.

En suma, no cambia la misión fundamental de la medicina, curar cuando se pueda, aliviar a menudo, consolar siempre, y hacerlo con humanidad y ética. La IA es una nueva herramienta al servicio de esa misión; no debe redirigir la misión hacia fines extraños. El profesionalismo consiste precisamente en salvaguardar que las innovaciones se integren sin traicionar esos ideales.

19. Bibliografía

1. Cruess RL, Cruess SR, Steinert Y. Medicine as a profession: a social contract. Lancet. 2000;356(9224):156–159. doi:10.1016/S0140- 6736(00)02426-7

2. Sullivan WM. Medicine under threat: professionalism and professional identity. CMAJ. 2020;192(6):E131–E132. doi:10.1503/cmaj.191638

3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. doi:10.1038/s41591-018-0300-7

4. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays. Proc Natl Acad Sci USA. 2017;114(16):4176–4181. doi:10.1073/pnas.1719072114

5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–118. doi:10.1038/nature21056

6. Alsaleh A, Alotaibi R, Alshehri MD, et al. A survey of explainable artificial intelligence in healthcare: concepts, applications, and challenges. Information Sciences. 2024;654:120126. doi:10.1016/j.ins.2024.120126

7. Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas sobre inteligencia artificial (Ley de Inteligencia Artificial). Diario Oficial de la Unión Europea, L 2024/1689, 12 de julio de 2024. Disponible en: https://eur-lex.europa.eu/legal content/ES/TXT/?uri=CELEX:32024R1689

8. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018;378(11):981– 983. doi:10.1056/NEJMp1714229

9. Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. Hum Factors. 1997;39(2):230– 253.doi:10.1518/001872097778543886

10.Goddard K, Roudsari A, Wyatt JC. Automation bias in decision support systems. J Am Med Inform Assoc. 2012;19(1):113–120. doi:10.1136/amiajnl-2011-000089

11.Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med. 2018;15(11):e1002689. doi:10.1371/journal.pmed.1002689

12.Emanuel EJ, Wachter RM. Artificial intelligence in health care. JAMA. 2019;321(2):127–128. doi:10.1001/jama.2018.18405

13.McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94. doi:10.1038/s41586-019-1799-6

14.Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019;103(2):167–175. doi:10.1136/bjophthalmol-2018-313173

15.Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice. Ann Intern Med. 2016;165(11):753–760. doi:10.7326/M16-0961

16.Shanafelt TD, Hasan O, Dyrbye LN, et al. Changes in burnout and satisfaction. Mayo Clin Proc. 2016;91(7):836–848. doi:10.1016/j.mayocp.2016.04.022

17.Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res. 2019;46(4):629–650. doi:10.1093/jcr/ucz013

18.World Health Organization. Ethics and governance of artificial intelligence for health. Geneva: WHO; 2021. Disponible en: https://www.who.int/publications/i/item/9789240029200

19.UNESCO. Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO; 2021. Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000380455

20.World Medical Association. Statement on Artificial Intelligence in Medicine. Ferney-Voltaire: WMA; 2025. Disponible en: https://www.wma.

net/policies-post/wma-statement-on-artificial-and augmented-intelligence-in-medical-care/

21.Masters K. Artificial intelligence in medical education. Med Teach. 2019;41(9):976–980. doi:10.1080/0142159X.2019.1595557

22.Park SH, Han K. Methodologic guide for evaluating clinical performance of AI. Radiology. 2020;294(1):13–20. doi:10.1148/radiol.2019191291

23.Price WN, Cohen IG. Liability for medical artificial intelligence. Harv Law Rev. 2019;132:1523–1569.

24.Gerke S, Minssen T, Cohen G. Ethical and legal challenges of AI in healthcare. Nat Med. 2020;26(1):105–107. doi:10.1038/s41591-019-0734-6

25.Topol EJ. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York: Basic Books; 2019.

26.Verghese A, Shah NH, Harrington RA. What this computer needs is a physician. N Engl J Med. 2018;378(22):2139–2141. doi:10.1056/ NEJMp 1802682

27.Consejo General de Colegios Oficiales de Médicos. Código de Deontología Médica: Guía de Ética Médica. Madrid: Organización Médica Colegial de España; 2022

28.Trzeciak S, Mazzarelli A. Compassionomics: The Revolutionary Scientific Evidence That Caring Makes a Difference. Pensacola, FL: Studer Group; 2019

29.Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023;183(6):589–596. doi:10.1001/jamainternmed.2023.1838

30.Hernández Gil ÁL, Moya García MI, Pérez Sarabia M, Díaz García J, Pérez Chica G, Miralles Linares F, et al. Declaración de Jaén sobre la medicina 5.0. Actual Med. 2025;110(821):51-3. doi:10.15568/am.2025.821.ds01

31.Martín-Quirós A, Jaén Cañadas M, Miralles Linares F. Structured prompting enhances ChatGPT-4.0 performance in ED triage: A complementary perspective. Am J Emerg Med. 2024. doi: 10.1016/j.ajem.2024.11.028

32.Martín-Quirós A, Jaen Cañadas M, Miralles Atencia P, Gómez Carrillo V, Ávila Huertas L, Miralles Linares F. Simulación de un flujo de trabajo asistido por inteligencia artificial en el triaje de un servicio de urgencias: estudio comparativo de eficiencia. Emergencias. 2025;37:474-80. doi:10.55633/s3me/086.2025

33.Salcedo-Lobera E, Ruiz-Esteban P, Miralles-Linares F. ¿Es posible el uso de la inteligencia artificial en las unidades de broncoscopia? Rev Pat Resp. 2025;28(3):167-9. doi:10.24875/RPR.24000037

34.Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2(3):158-64

35.Dembrower K, Crippa A, Colón E, Eklund M, Strand F. Artificial intelligence for breast cancer detection in screening mammography: a population based comparison of strategies. Lancet Digit Health. 2023;5(10):e703- e711. doi:10.1016/S2589-7500(23)00153-X

36.Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969

37.Reddy S, Allan S, Coghlan S, Cooper P. Establishing organizational AI governance in healthcare: a case study in Canada. npj Digit Med. 2025;8:522. doi:10.1038/s41746-025-01909-3

6

Ficciones fiduciarias y relación médico-paciente en la era del aprendizaje automático

Prof. Dr. Luis E. Echarte Alonso
Profesor Titular de la Unidad de Humanidades y Ética Médica.
Coordinador del Grupo de Investigación SOPHROSYNE: sobre salud y tecnologías emergentes.
Facultad de Medicina. Universidad de Navarra.

Resumen ejecutivo:

La rápida incorporación de los sistemas de aprendizaje automático en el ámbito sanitario está reconfigurando la relación médico-paciente. Esta transformación afecta no solo a su dimensión técnica, sino también al propio ethos médico. El presente texto analiza los principales factores que están impulsando esta profunda transformación, en la que la IA se revela como una nueva y poderosa cámara de eco de la deriva tecno-medicalizante, que aquí caracterizamos, en primer lugar, por la hiperobjetivación; en segundo lugar, por la hiperespecialización; y, finalmente, como estación término, por las ficciones fiduciarias.

En segundo lugar, en este análisis se identifican tres grandes tipos de sesgos asociados a la IA, concediendo especial relevancia al tercero: el sesgo tecnológico inherente, particularmente frecuente en periodos históricos como el actual, marcados por desarrollos tecnológicos abruptos. En este contexto se analizan cuatro conceptos marco fundamentales en medicina –vida, salud, autonomía e intersubjetividad– que están siendo redefinidos y, con ellos, también las dinámicas en la relación médico-paciente.

En tercer lugar, a este último sesgo se vinculan cuatro grandes catalizadores responsables del debilitamiento de la evaluación contextual y experiencial de la asistencia clínica: el incremento de la mediación en la relación interpersonal, el aumento de la complejidad tecnológica, el creciente predominio de métodos reductivos de conocimiento y la intensificación de intereses no terapéuticos. Además, entre las inercias que configuran la deriva señalada, se concede especial relevancia al ciclo de la metáfora computacional, entendido como el conjunto de creencias y, sobre todo, de prácticas asociadas a la delegación progresiva en la IA de tareas asistenciales. Este fenómeno se ve favorecido, en particular, tanto por la eficacia creciente de los automatismos en las tareas técnicas como por su capacidad para generar simulaciones afectivas.

Finalmente, se argumenta que este escenario hace ineludible el debate sobre la posibilidad de enseñar ética a las máquinas para evitar que deterioren la relación médico-paciente. Analizamos esta cuestión comparando, en particular, su eventual implementación desde el modelo principialista y desde el modelo profesionalista, para concluir que tal proyecto introduce en ambas aproximaciones riesgos inéditos, entre los cuales destacan el

espejismo de la antropomorfización de la máquina y su reverso complementario: la tecnificación de lo humano.

En las conclusiones finales se sostiene que la respuesta más adecuada a los riesgos asociados al cambio de paradigma médico no consiste tanto en programar máquinas morales como en mejorar la formación de los profesionales. La solución no puede limitarse al diseño de marcos regulatorios destinados a controlar la biotecnología, pues, si el diagnóstico aquí propuesto es correcto, tales normativas acabarán revelándose insuficientes o incluso perdiendo vigencia. Han de ir acompañadas de un proyecto formativo sólido, orientado tanto al estudio de los límites epistemológicos e instrumentales de la IA como a la adquisición de competencias interdisciplinares capaces de contrarrestar la fragmentación de la profesión y el proceso de tecnificación que la alimenta. En definitiva, se propone fortalecer la impronta universitaria como antídoto frente a la tentación de creer que el trabajo médico ha de ir orientándose exclusivamente hacia aquellas actividades que no puedan ser automatizadas.

Palabras clave:

Relación médico-paciente; ética de la inteligencia artificial; tecnificación médica; hiperespecialismo; ficciones fiduciarias; educación médica; vocación profesional.

Executive summary:

The rapid incorporation of machine-learning systems into healthcare is reshaping the doctor–patient relationship. This transformation affects not only its technical dimension but also the very ethos of medical practice. This text analyzes the main drivers of this profound shift, in which AI emerges as a new and powerful echo chamber for a techno-medicalizing drift, characterized here, first, by hyper-objectification; second, by hyper-specialization; and, ultimately, by fiduciary fictions.

Second, the analysis identifies three major types of AI-related bias, placing special emphasis on the third: inherent technological bias, particularly prevalent in historical periods like the present, marked by abrupt techno-

logical developments. In this context, four fundamental framing concepts in medicine—life, health, autonomy, and intersubjectivity—are examined as they are being redefined, along with the dynamics of the doctor–patient relationship.

Third, this latter bias is linked to four major catalysts responsible for weakening the contextual and experiential evaluation of clinical care: increased mediation within interpersonal relationships, growing technological complexity, the rising predominance of reductive methods of knowledge, and the intensification of non-therapeutic interests. In addition, among the inertias shaping the drift described, special relevance is given to the cycle of the computational metaphor, understood as the set of beliefs—and above all practices—associated with the progressive delegation of clinical tasks to AI. This phenomenon is facilitated, in particular, both by the increasing effectiveness of automatisms in technical tasks and by their ability to generate affective simulations.

Finally, it is argued that this scenario makes unavoidable the debate on whether it is possible to teach ethics to machines in order to prevent deterioration of the doctor–patient relationship. We analyze this issue by comparing, in particular, its potential implementation from the principlist model and from the professionalist model, concluding that such a project introduces unprecedented risks in both approaches—most notably the mirage of anthropomorphizing the machine and its complementary reverse: the technification of the human.

In the final conclusions, it is maintained that the most appropriate response to the risks associated with the shift in the medical paradigm lies less in programming moral machines than in improving the training of professionals. The solution cannot be confined to designing regulatory frameworks to control biotechnology, because, if the diagnosis proposed here is correct, such norms will ultimately prove insufficient or even lose relevance. They must be accompanied by a robust educational project, oriented both toward studying the epistemological and instrumental limits of AI and toward acquiring interdisciplinary competencies capable of counteracting the fragmentation of the profession and the process of technification that fuels it. In short, the proposal is to strengthen the university imprint as an antidote to the temptation to believe that medical work should progressively be oriented exclusively toward those activities that cannot be automated.

6 Ficciones fiduciarias y relación médico-paciente en la era del aprendizaje automático

Keywords:

doctor–patient relationship; AI ethics; medical technification; hyper-specialization; fiduciary fictions; medical education; professional vocation.

Ideas fuerza:

- La IA no solo introduce nuevas herramientas: reorienta el ethos clínico y la relación médico-paciente hacia una deriva tecno-medicalizante que progresa desde **la hiperobjetivación y la hiperespecialización hasta las ficciones fiduciarias.**

- Los sesgos de la IA no se reducen al "mal dato": hay **sesgos deliberados** (diseño/mercado), **sesgos emergentes** (proxies y desigualdad estructural) y, sobre todo, un **sesgo tecnológico inherente** que altera qué cuenta como conocimiento y qué fines terminan priorizándose.

- Cuatro conceptos-marco (vida, salud, autonomía e intersubjetividad) se están redefiniendo bajo lógicas de predicción, control y estandarización; el riesgo es confundir lo cuantificable con lo real y empujar una "medicalización de la felicidad".

- La pérdida de evaluación clínica contextual se acelera por cuatro catalizadores: más mediación tecnológica, mayor complejidad, predominio de métodos reductivos e intensificación de intereses no terapéuticos; el "ciclo de la metáfora computacional" favorece delegación, descualificación y la confusión entre autonomía y automatismo.

- "Enseñar ética a máquinas" (desde el principialismo o el profesionalismo) introduce riesgos inéditos: el espejismo de antropomorfizar la IA y su reverso, la tecnificación de lo humano; la salida prioritaria es **mejorar la formación** (límites epistemológicos/instrumentales + competencias interdisciplinares), más que programar "máquinas morales".

Key messages:

- AI is not merely introducing new tools: it is reshaping clinical ethos and the doctor–patient relationship through a techno-medicalizing drift that moves from **hyper-objectification and hyper-specialization toward fiduciary fictions.**

- AI bias is not just "bad data": there are deliberate biases (design/market), **emergent biases** (proxies and structural inequality), and—above all—an **inherent technological bias** that changes what counts as knowledge and which ends become prioritized.

- Four framing concepts (life, health, autonomy, and intersubjectivity) are being redefined under logics of prediction, control, and standardization; the risk is mistaking what is quantifiable for what is real and driving a "medicalization of happiness."

- The erosion of contextual clinical judgment is accelerated by four catalysts: greater technological mediation, increasing complexity, the dominance of reductive methods, and intensifying non-therapeutic interests; the "cycle of the computational metaphor" promotes delegation, deskilling, and the confusion of autonomy with automatism.

- "Teaching ethics to machines" (via principlism or professionalism) introduces unprecedented risks: the mirage of anthropomorphizing AI and its reverse, the technification of the human; the priority response is **better professional education** (epistemic/instrumental limits + interdisciplinary competencies), rather than programming "moral machines."

6 Ficciones fiduciarias y relación médico-paciente en la era del aprendizaje automático

Sumario:

1. Introducción. A los fines por los medios

La integración de la inteligencia artificial (IA) en los sistemas de salud está llamada a introducir transformaciones relevantes en la práctica médica contemporánea. Sin embargo, la literatura especializada muestra dos posiciones claramente diferenciadas ante este cambio. Por un lado, algunos autores sostienen que la automatización de los procesos clínicos irá acompañada de una modificación sustancial del rol del profesional sanitario, hasta el punto de reconfigurar la relación tradicional entre médico y paciente (Adams 2025; Buhr et al 2025; Verghese et al 2018; Floridi 2018). Por otro lado, existen posiciones que defienden que, si bien estas tecnologías modificarán las cuestiones organizativas, técnicas o de eficiencia, así como las formas de prestación de la atención sanitaria, estos cambios no alterarán los fines fundamentales que caracterizan la relación médico-paciente (Mittelstadt 2022, 41-43; London 2020; Reddy et al 2020).

El presente trabajo se inscribe en el debate clásico sobre la relación entre fines y medios. Cabe adelantar sus conclusiones: si bien la inteligencia artificial (IA) encierra un enorme potencial transformador en el ámbito biosanitario con respecto a los medios clínicos y de investigación, plantea asimismo riesgos relevantes, entre los más preocupantes, los que conciernen a la configuración de los fines de la práctica profesional. A desarrollar y fundamentar esta afirmación se dedica lo que sigue.

2. Tres niveles de afectación tecnológica

Para la Organización Mundial de la Salud, el propósito fundamental de la medicina consiste en trabajar por la salud y el cuidado del paciente sin olvidar la sociedad que lo integra y lo ampara (OMS, 2014). Estos fines se alcanzan gracias al buen entendimiento entre el médico y el paciente que, a su vez, se cimienta en un marco de reconocimiento recíproco. Dicha relación no se limita al plano intelectual o técnico, sino que se construye también sobre valores compartidos como la dignidad humana, el derecho a la privacidad, el consentimiento informado, entre otros (GMC, 2024).

Uno de los rasgos esenciales de la relación médico-paciente es su asimetría intrínseca: el paciente se encuentra en una posición de particular vulnerabilidad, como sujeto doliente con la autonomía comprometida. Esta necesidad de ayuda, desde la perspectiva médica, toma forma de deber fiduciario: el profesional no utiliza su posición privilegiada para abusar de la confianza que el paciente deposita en él (Pellegrino y Thomasma, 1993, pp. 65-67).

Respecto a los medios, pueden distinguirse tres grandes niveles. En primer lugar, el juicio clínico experto, es decir, el uso del conocimiento científico y el razonamiento diagnóstico que el médico adquiere tanto en su formación académica como en su experiencia profesional. Este primer medio, irreductible a procedimientos puramente técnicos o a reglas algorítmicas, permite al médico ejercer la medicina considerando a cada paciente como un individuo único. En segundo lugar, se encuentran las tecnologías médicas y los sistemas técnicos orientados al diagnóstico y tratamiento, en los que hay que incluir también los dispositivos de monitorización, el software clínico de apoyo a la decisión o las plataformas de historia clínica electrónica. Por último, la práctica médica también se sustenta en normas institucionales, contextos organizativos y procedimientos, tales como protocolos, guías clínicas, estructuras hospitalarias o sistemas de gestión de datos. Aunque estos elementos no son “herramientas” en un sentido técnico estricto, configuran profundamente la práctica médica al definir los marcos en los que se toman decisiones, lo cual influye directamente en los valores que se priorizan, como la eficiencia, la rapidez, la seguridad, la empatía o la autonomía.

La introducción de sistemas de inteligencia artificial (IA) en el ámbito médico está impactando, de forma simultánea, pero todavía diferenciada en estos tres niveles clave. Este efecto puede observarse hoy, por ejemplo, en apli-

caciones para la formación médica, para la interpretación automatizada de imágenes diagnósticas o para la reorganización de los flujos del trabajo asistencial. Es precisamente esta triple capacidad de intervención y, en especial, su potencial para integrarse dentro de un mismo plan estratégico lo que lleva a Peltonen, Topaz y Zhang (2025) a afirmar que la inteligencia artificial transformará cualitativamente –y en un futuro cercano– la práctica clínica contemporánea no solo con respecto a los medios. Dado que la IA es una herramienta técnica con capacidad para reorganizar, de forma simultánea, la cognición clínica, la acción técnica y el marco institucional, también puede convertirse en un factor estructurante que influya no solo en cómo se priorizan y alcanzan los fines de la medicina en la práctica clínica, sino también en la redefinición de dichos fines.

2.1 El supuesto de neutralidad

La cuestión de cómo la inteligencia artificial (IA) afecta a los fines de la medicina forma parte de un debate más amplio sobre la supuesta neutralidad de la tecnología. Y hoy es más evidente que nunca que el valor de los instrumentos tecnológicos no depende únicamente del uso que se haga de ellos. Esta afirmación puede sostenerse, al menos, en tres sentidos distintos.

El primero se refiere a cómo el diseño de un instrumento puede estar intencionalmente sesgado. Un caso muy conocido es el Proyecto IBM Watson Health para cáncer. En el desarrollo del software Watson for Oncology, se eligió entrenar el sistema en IA casi exclusivamente con datos del Memorial Sloan Kettering Cancer Center, un hospital privado de élite con población mayoritariamente blanca. Los expertos de IBM alertaron internamente que el sistema fallaría con poblaciones más diversas (inmigrantes, afroamericanos, comunidades rurales), pero la dirección decidió ignorar esas advertencias, alegando que introducir más variedad “complicaría el modelo”. El sesgo fue una decisión estratégica consciente para acelerar la salida al mercado del producto y cerrar acuerdos comerciales con hospitales similares al centro de entrenamiento (Rigby 2019; Frank 2019). Casos como éste han motivado la proliferación de marcos regulatorios orientados al “diseño responsable” de sistemas de inteligencia artificial (IA), es decir, herramientas concebidas para facilitar la incorporación de principios como la equidad, la explicabilidad y la no discriminación desde las etapas iniciales del desarrollo tecnológico.

En un contexto donde las presiones no asistenciales crecen y la capacidad del médico para evaluar la neutralidad del software es cada vez más limitada,

el establecimiento de nuevas medidas éticas y legales se vuelve crucial, no solo para disipar las dudas en torno a la equidad de las decisiones clínicas automatizadas, sino, más importante aún, para evitar que los fines de la medicina se desnaturalicen. Porque existe el riesgo de que dichos fines acaben reducidos a meras apariencias que encubran intereses ajenos a la razón de ser de las prácticas del cuidado, intereses que, con creciente frecuencia, operan sin que el propio profesional sea plenamente consciente de su influencia.

La IA revela, además, un segundo modo específico en el que la supuesta neutralidad tecnológica queda en entredicho. En estos sistemas, una parte de la secuencia algorítmica es "aprendida": la IA es capaz de ajustar sus parámetros internos en función de los datos de entrada con el fin de optimizar su rendimiento. En el caso del denominado aprendizaje no supervisado, el algoritmo no solo ejecuta instrucciones, sino determina por sí mismo las reglas de toma de decisiones. En otras palabras, no se limita a facilitar el cumplimiento de metas previamente definidas, pues puede llegar a reformular la forma en que esos fines son comprendidos, priorizados o incluso sustituidos. Al introducir nuevas lógicas operativas –como la eficiencia, la predicción, la estandarización o la automatización–, la IA es capaz de desplazar, de forma no intencionada por sus programadores humanos, valores centrales como la singularidad del paciente, la deliberación ética o la relación interpersonal. En consecuencia, deja de ser un mero medio técnico neutral para convertirse en un actor estructurante que reconfigura los marcos en los que se toman decisiones clínicas y se define qué cuenta como buena práctica médica (Amann et al., 2020).

2.2 Copilotaje tecnológico

La afirmación de que un programa informático tiene la capacidad de aprender resulta controvertida si tomamos el término en su sentido fuerte, es decir, como un proceso que implica comprensión, experiencia subjetiva y adaptación contextual significativa. Desde esta perspectiva, lo que comúnmente se denomina aprendizaje automático no constituiría un aprendizaje en sentido humano, sino una forma de optimización estadística de patrones dentro de los límites predefinidos por los programadores y los datos suministrados. Sin embargo, en un sentido más débil –y en la medida en que esos límites pueden llegar a desplazarse– resulta innegable que esta tecnología ha comenzado a desempeñar tareas que, hasta hace poco, eran exclusivas de agentes considerados inteligentes (Bertoncini y Serafim

2023). Por eso, mejor que concebir la inteligencia artificial (IA) como una herramienta completamente subordinada, resulta más adecuado entenderla como una suerte de copiloto: no dirige la nave, pero actúa como segundo de a bordo, es decir, no se limita a ejecutar órdenes, sino que es capaz de cierta novedad también en lo que respecta a los fines. De ahí el amplio abanico de posibilidades y oportunidades que esta tecnología ofrece, pero también los nuevos riesgos que conlleva.

Uno de los fallos de desplazamiento no intencionado que más eco ha recibido es el caso Optum, relacionado con un algoritmo de gestión poblacional desarrollado por una empresa norteamericana. En su diseño, orientado a identificar a los pacientes que requerían intervenciones médicas intensivas, el sistema utilizaba el coste sanitario como variable proxy para medir la necesidad clínica. Este criterio, aunque aparentemente neutro, condujo a la exclusión sistemática de pacientes afroamericanos debido a las desigualdades históricas en el acceso a la atención sanitaria. Como consecuencia, el algoritmo asignaba menos recursos a quienes más los necesitaban, perpetuando una desigualdad que sus programadores ni habían previsto ni seguramente habrían deseado (Obermeyer et al., 2019).

El escándalo de Optum reveló que, a diferencia de los sesgos introducidos de forma deliberada, los sesgos no intencionados resultan mucho más difíciles de identificar y corregir. En estos casos, el problema suele emerger de la interacción entre varios factores: conjuntos de datos de entrenamiento incompletos o desbalanceados (una de las clases domina sobre las demás), la elección inadecuada de variables proxy (como utilizar el coste económico en lugar de la necesidad clínica), o la reproducción de patrones históricos de desigualdad ya presentes en el sistema sanitario. Estos elementos suelen quedar ocultos dentro del funcionamiento estadístico del algoritmo, que puede ofrecer resultados aparentemente precisos y coherentes sin levantar sospechas inmediatas (Lumbreras et al 2025). El sesgo no se encuentra en una instrucción previa del programador, ni tampoco en una línea de código concreta, sino en la lógica relacional aprendida por el sistema.

2.3 Instrumentos racionales

La inteligencia artificial puede incidir en los fines de la medicina de un tercer modo, igualmente no intencionado –como el segundo–, aunque más abstracto y, por ello, más difícil de advertir. Comprender la naturaleza de este sesgo, al que llamaré sesgo tecnológico inherente, exige examinar primero

los fundamentos mismos de la tecnología. Para ello, comenzaremos por una clarificación terminológica.

La diferencia entre los conceptos de instrumento y herramienta es accidental: ambos se refieren a medios utilizados para alcanzar un fin, aunque el segundo suele aplicarse a objetos físicos y extracorpóreos. En este sentido –y solo en este sentido– encontramos en el mundo animal numerosas especies que utilizan herramientas. Por ejemplo, los chimpancés emplean ramas para atrapar insectos o usan piedras para romper nueces. Sin embargo, esta habilidad presenta una limitación importante: las herramientas son utilizadas para unos pocos propósitos y, con frecuencia, en un mismo contexto.

Las herramientas en el mundo animal carecen de la versatilidad de las humanas o, mejor dicho, los seres humanos son capaces de descubrir nuevos modos de utilizar las herramientas, ya para cumplir con un mismo fin o, lo que es aún más relevante, ya para cumplir con fines distintos. Esta capacidad –fruto de la singular inteligencia creativa del ser humano, de su racionalidad– constituye el fundamento de eso a lo que damos el nombre de tecnología. Las herramientas tecnológicas no deben entenderse como un simple conjunto de instrumentos, por diversos que sean, sino como el resultado de la aplicación práctica de un sistema articulado de saberes teóricos. Solo así parece justificado que el ser humano pueda mirar un instrumento más allá de lo que es, más allá del fin para el que fue concebido (Heersmink 2022). Y en efecto, la historia de la medicina está repleta de ejemplos de instrumentos, fármacos y técnicas que acabaron teniendo un uso muy distinto al originalmente previsto. Véase, como ejemplo reciente, el caso de materiales quirúrgicos clásicos como el ácido poliláctico, que primeramente era utilizado en suturas internas y hoy se emplea en cultivos tridimensionales de células madre, actuando como andamios (scaffolds) para la regeneración de órganos.

La clave que nos interesa en esta cuestión es que la capacidad de innovación tecnológica no parece depender exclusivamente del sujeto cognoscente, sino también de los horizontes de posibilidad que cada creación instrumental abre. Como ya anticipó el filósofo estadounidense John Dewey, la inteligencia práctica no consiste simplemente en aplicar fines a medios preexistentes, sino en mantener una relación dinámica con el entorno. Los instrumentos –ya sean formales o físicos– deben entenderse como estructuras activas, capaces de reconfigurar la manera en que se comprenden y abordan los problemas (Dewey 1938, 391-392). Así, toda creación técnica

no responde únicamente a una necesidad previa; en ocasiones, y dependiendo de su potencia instrumental, es la propia tecnología la que genera nuevas necesidades. De ahí que la función epistemológica de la tecnología constituya un tercer argumento para cuestionar la supuesta neutralidad instrumental, especialmente en el caso de tecnologías como la inteligencia artificial, pues –como fue indicado al inicio de la segunda sección– ejerce un impacto transformador de amplio alcance.

No hace falta adscribirse al pragmatismo epistemológico de Dewey para reconocer que ciertos fines –y también algunas teorías sobre la realidad– no solo han sido propiciados por los medios disponibles para conocer y manipular el entorno, sino que dependen estrechamente de ellos. En consecuencia, su valor ontológico resulta limitado: dicen más sobre nuestras herramientas que sobre la naturaleza de las cosas en sí. Parece razonable pensar, además, que este fenómeno se haya visto intensificado en periodos históricos marcados por desarrollos tecnológicos abruptos. A esto debemos sumar, además, que la innovación esté hoy tan estrechamente vinculada a la irrupción de los nuevos copilotos tecnológicos. En definitiva, ambos factores contribuyen al proceso de desontologización, a menudo no intencional, que están sufriendo tanto las ciencias descriptivas como las normativas.

2.4 El ciclo de la metáfora computacional

Que la tecnología sea sustancialmente parcial, o lo que es lo mismo, que introduzca sesgos en nuestra manera de enfrentarnos al mundo, es una idea bien conocida por cualquier investigador. La ciencia moderna debe buena parte de su éxito a las aproximaciones analíticas y, por tanto, al uso de métodos reductivos para el estudio de los fenómenos. Ahora bien, una buena formación científica implica, entre otras habilidades, aprender a evitar que esta reducción –o simplificación– metodológica conduzca al científico a caer en el espejismo del reduccionismo ontológico: creer que lo que el instrumental mide es lo único que puede ser conocido o, peor aún, lo único real. El conocimiento de los límites de los métodos aplicados forma parte fundamental del buen quehacer del investigador.

La tecnología potencia la innovación de un modo similar a como los métodos de conocimiento impulsan el desarrollo científico, a condición de que el investigador sea consciente y aprenda a superar la direccionalidad inevitable –el sesgo– que acompaña a ambos.

Este aprendizaje consciente constituye una de las líneas rojas que separan la inteligencia humana de la no humana. La racionalidad nos permite elevarnos por encima de las reglas del método o del instrumento y comprender la naturaleza misma del juego, ya sea teórico o práctico –no de un juego concreto, sino de la capacidad de entender cualquier juego. En esta observación se justifica la afirmación, en sentido fuerte, de que el ser humano es la única especie capaz de jugar (Hamayon 2016, xviii–xix). Jugar implica, entre otras cosas, distinguir el juicio sobre la realidad (por ejemplo, acerca de la verdad o del bien) de la realidad misma, así como experimentar el gozo contemplativo que acontece en dicha iluminación. En contrapartida, aunque la inteligencia artificial pueda cambiar las reglas del juego –como se ha señalado, modificar medios e incluso fines–, permanece ajena tanto al sentido del juego como al de sus transformaciones. El problema es que esta incapacidad ontológica no resulta fácilmente identificable en términos de eficiencia o productividad, lo que, como vamos a ver más tarde –al comparar la simulación y la réplica de la inteligencia natural– genera importantes problemas prácticos.

Tres son los grandes inconvenientes de ignorar la falacia de la neutralidad tecnológica en el tercer sentido aquí planteado. El primero, y quizá el más evidente, es la adquisición de una comprensión distorsionada de la realidad, problema que conlleva, de forma inevitable, modos erróneos de interactuar con ella. El segundo se refiere a la merma de las capacidades para generar verdadera innovación a partir de los escenarios emergentes. La innovación que puede generar una inteligencia artificial, si Hamayon tiene razón sobre el juego humano, sería cualitativamente inferior, aunque no lo pareciera en el corto o medio plazo. El tercero alude a cómo la creciente identificación entre el pensamiento humano y el funcionamiento de la máquina –la hipótesis de la metáfora computacional de la mente–, acabará derivando, no solo en una progresiva desfiguración de la autocomprensión humana, sino también, en la práctica, en que se delegue en la inteligencia artificial competencias y responsabilidades que exceden las capacidades de cualquier algoritmo. Este tercer error amplificaría y normalizaría los dos anteriores, cerrando así un círculo vicioso que, por motivos obvios, resultará difícil revertir una vez arraigado. Desarrollemos a continuación esta idea en el contexto asistencial.

3. La deriva tecno-medicalizante

Cuatro de los conceptos fundamentales para el sentido de la actividad médica –su ethos– corren el riesgo de verse alterados con la introducción de la IA: la noción de vida, de salud, de autonomía y de intersubjetividad.

Pocas decisiones ilustran con tanta claridad la falacia de la neutralidad tecnológica como la aceptación médica de los criterios de muerte encefálica. El trasfondo de esta cuestión remite a los progresos en materia de trasplante de órganos y soporte vital que tuvieron lugar a lo largo de las décadas de 1950 y 1960. La innovación tecnológica propició el debate, pero cabe preguntarse si también llegó a condicionarlo. Sin duda, el informe de Harvard de 1968 suscitó numerosas sospechas: ¿la redefinición de la muerte respondía a un intento de describir el cese de la vida a la luz de los nuevos hallazgos científicos, o a la necesidad de justificar, desde un punto de vista operativo, la extracción de órganos viables lo antes posible? (Sulmasy et al 2024; Gardner 2019). Se trata, sin duda, de un tema controvertido en torno al riesgo de medicalización de la muerte. De nuevo, ¿lo que antes era una realidad comprensible en el amplio marco de la experiencia humana puede haberse transformado en una categoría marcada por su facticidad, esto es, reducida ontológicamente a criterios objetivos de predicción y control? (Donley & Fannin 2024).

No es una cuestión ajena a la inteligencia artificial. Como se ha dicho, en la medida en que sea ignorado el límite metodológico-instrumental, es inevitable que este tipo de programas hagan de cámaras de eco y acentúen la brecha entre la percepción humana y el diagnóstico clínico (García Abejas et al 2025; De Panfilis et al 2025). Por ejemplo, en el corto plazo, la IA podría reforzar una visión excesivamente objetivada del morir, algo especialmente relevante en entornos críticos como las unidades de cuidados intensivos o las unidades de trasplantes. En situaciones de incertidumbre clínica, su copilotaje podría terminar acelerando los desenlaces irreversibles. Por otro lado, precisamente por la naturaleza sensible de este tipo de decisiones, los profesionales podrían desear delegarlas en dichos programas automatizados. A largo plazo, cabría temer que la inteligencia artificial pudiera favorecer la redefinición misma de la concepción de vida humana: de entenderla como una realidad dotada de sentido en sí misma, a tomarla como un fenómeno estrictamente funcional, gestionable según parámetros utilitaristas e intereses comerciales.

Entonces, ¿cómo enseñar a una máquina que se pueden aceptar los criterios de muerte cerebral y, al mismo tiempo, sostener que estos no necesariamente sirven para comprender qué es la vida o qué significa morir? No será una tarea sencilla mientras que el sistema sea incapaz de elevarse por encima del juego algorítmico para el que ha sido programado.

3.1 Catalizadores de la hiperobjetivación

Adentrémonos en ese escenario del largo plazo, menos lejano de lo que parece. Porque vida y salud son conceptos íntimamente entrelazados. La salud actúa, dentro del horizonte más amplio de la vida, como criterio orientador de la acción clínica en su dimensión inmediata y operativa.

En la relación médico-paciente, el facultativo determina el estado de salud del enfermo mediante una evaluación contextual y experiencial, donde la aproximación objetiva y subjetiva quedan integradas en una misma comprensión unitaria de la situación clínica. Sin embargo, como advierte Mittelstadt, esta comprensión se ve amenazada cuando tecnologías de monitorización remota u otros sistemas de recopilación de datos en entornos no presenciales adquieren protagonismo. “Las representaciones de datos de los pacientes pueden llegar a considerarse una medida «objetiva» de la salud y el bienestar, lo que reduce la importancia de los factores contextuales de la salud o la visión del paciente como persona socialmente integrada.” (Mittelstadt 2022, 51).

La falsa creencia a la que hace referencia Mittelstadt se ve reforzada por cuatro grandes catalizadores. El primero tiene que ver con cómo el uso de dicha tecnología, aun cuando la evaluación clínica sea incluso presencial, puede generar que el médico se acabe acostumbrando a reemplazar la mirada directa y totalizante hacia el paciente por una imagen mediada y parcial, que, sin embargo, ofrece la promesa de recomendaciones novedosas y personalizadas. El coste de este nuevo modo de mirar al paciente no es solo epistémico, pues el progresivo aumento de la mediación tecnológica de la relación clínica puede debilitar el reconocimiento compasivo del individuo sufriente al que precisamente se busca ayudar (Čartolovni 2023; Verghese 2018). En la sección 3.4 se desarrollará más este segundo asunto.

Un segundo catalizador es el aura que suele rodear a las tecnologías más complejas, que induce a esperar de ellas más de lo que realmente pueden ofrecer. Este fenómeno se intensifica especialmente en contextos de auge

tecnológico como el que actualmente rodea a la inteligencia artificial, cuya popularidad y promesas posiblemente terminen por moderarse con el tiempo. El problema está en que mientras tanto, conducen la práctica asistencial y la investigación biomédica por senderos indeseados e incluso irreversibles (De Togni et al 2024, Echarte 2023a). Igual que hay falsas expectativas que, a pesar de todo, favorecen el progreso, hay también otras que no hacen sino entorpecerlo.

El tercer catalizador está también relacionado con la complejidad, aunque ahora sobre el objeto de estudio y manipulación. En campos todavía tan enigmáticos como puede ser la psiquiatría o en especialidades con metas amplias y difusas como la prevención y la promoción de la salud y la calidad de vida, la tentación de privilegiar los métodos más reductivos viene justificada por la aparente solidez de lo cuantificable (Gaud et al 2024; Santos y Nazaré 2025). Si además la tecnología promete hacer avanzar la investigación –como en parte, ocurrirá–, la presión por adoptar enfoques reductivos se ve aún más reforzada, al quedar dichos enfoques asociados no solo con la eficacia operativa, sino también con la idea más general y poderosa de progreso científico. Por supuesto, todo ello a expensas del conocimiento contextual y tácito –no objetivable–, difícilmente modelizable, pero no por ello menos relevante.

Muchos proyectos actuales de explicabilidad computacional pasan por alto esta dimensión de lo real. Ello obedece a una limitación pragmática: los modelos de explicabilidad se diseñan para hacer inteligibles las correlaciones internas del sistema o para justificar decisiones en términos formalizables, no para reconstruir la hermenéutica de la práctica clínica. Desde esta perspectiva, la explicabilidad resulta insuficiente, pues opera dentro del mismo marco metodológico que privilegia lo cuantificable. En consecuencia, el problema no radica únicamente en que los modelos explicativos actuales sean técnicamente imperfectos, sino en que el propio ideal de explicación computacional es estructuralmente incapaz de captar determinadas dimensiones constitutivas de la práctica médica. Por esto, la controversia va mucho más allá del debate acerca de si deben emplearse exclusivamente algoritmos transparentes o del cálculo riesgo-beneficio que implicaría el uso de modelos de caja negra, generalmente dotados de mayor poder predictivo (Adams 2023; Ursin et al. 2023). Porque se hace preciso garantizar que los sistemas puedan ofrecer explicaciones formales de sus decisiones, en efecto, pero también preguntarse qué tipo de racionalidad se legitima implícitamente cuando se identifica la comprensión con la mera trazabilidad técnica. Un

algoritmo puede ser transparente en sus reglas o en la ponderación de sus parámetros estadísticos y, sin embargo, seguir operando dentro de un marco reductivo que excluya dimensiones esenciales del juicio clínico (Herzog 2022; Kannetzky 2002). Desgraciadamente, parece mayoritaria la posición de los especialistas en ética de la IA que centran sus esfuerzos en los requisitos de transparencia y control técnico, lo que favorece que el debate quede atrapado en cuestiones meramente instrumentales. Los árboles impiden ver el bosque; de ahí la necesidad de buscar modos de ascender hasta las copas. La segunda parte del capítulo examina con mayor detenimiento esta cuestión.

Pasando ya al último catalizador, paralelo a los intereses partidistas señalados en el debate sobre la muerte encefálica, se encuentra el elemento comercial. La cuestión aquí es que aquello que puede objetivarse es también más susceptible de convertirse en mercancía. Biomarcadores, algoritmos diagnósticos o escalas de riesgo y de calidad de vida pueden adquirir una legitimidad desproporcionada, facilitando la oferta de fármacos, dispositivos o programas cuyas aspiraciones comerciales eclipsan las terapéuticas, con las consecuencias que de esto se derivan.

Bajo dicha luz, debemos temer el riesgo de que la inteligencia artificial actúe también como cámara de eco dentro del fenómeno más amplio de la medicalización de la condición humana, que bien podríamos denominar medicalización de la felicidad: una felicidad reducida a salud, y una salud, a su vez, objetivada. Y en efecto, dos usos clínicos en los que la IA está demostrando mayor potencial son en la búsqueda del diagnóstico precoz y en la prevención. Las nuevas oleadas de indicadores clínicos que pueda generar la IA, sin una adecuada contextualización, generarán alarmas o, al menos, sobredimensionarán la salud sobre algo más general e importante que es el buen vivir.

La mayoría de los peligros mencionados en esta sección ya existían antes de la IA. Se estudia la medicalización como problema social desde la década de los setenta del siglo pasado (Conrad 2007). No es casual que una década más tarde, se cree el concepto de prevención cuaternaria con el que, dentro del contexto de la medicina general y la atención primaria, se pretenda proteger a los pacientes de intervenciones médicas innecesarias, evitar el sobrediagnóstico, la medicalización excesiva y el daño iatrogénico (Otte et al 2024). Como ya se apuntó al final de la sección anterior, uno de los grandes retos en computación consiste en hacer entender a las máquinas

que cuando la parte –aunque real– es asumida como el todo, se convierte en la peor de las ficciones: una vía muerta para el progreso científico y humano. Mientras siga siendo eso, solo un reto, hay razones para temer que la IA no solo acelere el proceso de medicalización, sino que lo oriente hacia una muy concreta dirección. Veamos a continuación algunos rasgos de esta nueva deriva tecno-medicalizante.

3.2 La Fase de Mediación Tecnológica

La autonomía, tercer concepto-marco en riesgo dentro del ámbito sanitario, requiere también una revisión de unos antecedentes que han facilitado la incorporación de la inteligencia artificial en nuestras vidas y además explican por qué dicha tecnología incide hoy de manera tan directa en el significado mismo de autonomía.

José Ramón Repullo, en un trabajo publicado en 2024, define la tecnomedicina como un cambio de equilibrio donde el componente instrumental y tecnológico acaba predominando sobre la interacción clínica interpersonal. En la misma línea de ideas que aquí se está presentando, atribuye al crecimiento exponencial del conocimiento y la innovación tecnológica la transformación de la práctica médica. Más específicamente, identifica tres fases o cambios en la evolución de la relación médico-paciente bajo esta influencia.

En la primera, la Fase de Relación Directa, el médico responsable entabla una relación médico-paciente directa con el paciente. En esta etapa, el equipo de especialistas y el "armario" de instrumentos tecnológicos están a disposición del médico responsable, actuando en el papel de interconsultores. En una segunda, la Fase de Dispersión o Paralelismo, desaparece la figura del médico responsable. El paciente establece relaciones en paralelo con diversos especialistas, lo que provoca que el proceso clínico pierda su integralidad y no se distinga el problema principal de las comorbilidades. Aquí, los especialistas comienzan ya a identificarse más con las técnicas que manejan que con el paciente mismo. Por último, en la Fase de Mediación Tecnológica, las tecnologías salen de "entre bastidores" para protagonizar la demanda directa del paciente. Los clínicos asumen el rol de "tecnólogos del conocimiento", limitándose a ejecutar técnicas, pero alejándose de la indicación de las mismas. En este punto, la relación médico-paciente ya no es inmediata, sino mediada por la tecnología.

En la deriva que describe Repullo, la tecnología gana protagonismo progresivamente, no solo como herramienta, sino como referente epistémico y organizativo. El médico –tradicionalmente piloto de la relación clínica– comienza a ocupar un lugar subsidiario, como ejecutor de indicaciones derivadas de sistemas, algoritmos o de las expectativas inducidas por el propio paciente, situación en la que, como señala Mittelstadt, "se corre el riesgo de reducir al médico a un mero proveedor de servicios, incapaz de ejercer toda la gama de virtudes médicas y normas internas de la práctica médica" (Mittelstadt 2022, 43). La autonomía del paciente se ve así progresivamente desplazada hacia un plano de decisión meramente reactivo ante las opciones generadas por un entorno tecnológicamente prescrito. En lugar de aumentar las posibilidades de deliberación conjunta, la tecnomedicina, tal como se perfila en esta trayectoria, tiende a colonizar el juicio clínico, reduciendo el margen de maniobra tanto del médico como del paciente –los que antes eran los verdaderos agentes morales de la relación asistencial.

Otro desafío para la autonomía que plantean las etapas más avanzadas de la Fase de Mediación Tecnológica es que, a medida que las herramientas se vuelven más complejas, resulta también más difícil para el médico comprender correctamente su funcionamiento y, por tanto, que pueda controlarlas adecuadamente (Matthias 2004). Esto incrementa la necesidad de confiar en ellas, lo que, a su vez, socava tanto la capacidad del profesional para detectar fallos como su habilidad para adaptar el uso de éstas a cada situación clínica –o incluso para innovar, ampliar el abanico de posibilidades del instrumento –su potencial versátil.

Por último, el aumento de la distancia herramienta-usuario se asocia a tres grandes riesgos adicionales relacionados con la autonomía. En primer lugar, la relación de confianza puede amplificarse por motivos injustificados, lo que necesariamente termina derivando en decisiones asistenciales erróneas. Por ejemplo, como señala Zarsky, los médicos pueden asumir las recomendaciones de un programa informático no por una eficacia clínica demostrada, sino por la percepción de su objetividad, precisión o complejidad (Zarsky 2016). En segundo lugar, el sentido de responsabilidad en la toma de estas decisiones copilotadas puede acabar diluido, desplazándose hacia una red de actores impersonales (Davis et al., 2013). La historia ofrece numerosos ejemplos de decisiones catastróficas ejecutadas en contextos marcados por este clima de conciencias atenuadas. Finalmente, la excesiva confianza en la tecnología y el debilitamiento de la conciencia moral aumentaría la delegación en la tecnología y, por ende, el riesgo de descualificación: término

para describir cómo el desarrollo de habilidades y virtudes se ve inhibido por el abandono del ejercicio de las prácticas (Coeckelbergh 2013). La paciencia se pierde cuando deja de practicarse. Como veremos a continuación, este último riesgo resulta especialmente relevante en el ciclo de la metáfora computacional.

3.3 Autonomía y automatismos

De entre los peligros derivados del debilitamiento de la autonomía médica mencionados arriba, Repullo destaca aquellos relacionados con la fragmentación y la atomización asistencial. Por un lado, el desajuste provocado por la participación no articulada de múltiples especialistas incrementa los costes y los riesgos para la seguridad del paciente, especialmente en casos crónicos o de alta fragilidad. Por otro lado, cuando los procedimientos y las tecnologías abandonan su papel instrumental y asumen el protagonismo, se instala progresivamente una confusión reductiva que puede llevar –implícita o explícitamente– a médicos y pacientes a identificar la autonomía con el automatismo.

La nueva inteligencia artificial representa, en este sentido, una vuelta de tuerca más en la doble deriva hacia el pilotaje tecnológico y hacia el espejismo de antropomorfización. La delegación del juicio ético constituiría, en ambos aspectos, el penúltimo escalón en la materialización final del reemplazo. Lejos de ser una delegación caprichosa, podemos definirla incluso como bien intencionada: la principal razón para crear máquinas éticas, en el contexto actual, consistiría en intentar resolver así el problema de fragmentación que, paradójicamente, estas mismas tecnologías parecen generar en la Fase de Mediación Tecnológica. Fijémonos en que el objetivo concreto de lograr que una máquina tome decisiones éticas en el contexto asistencial es hacerla capaz de una comprensión completa y unificada de la situación del paciente.

Pero ¿es posible programar una máquina para que tome decisiones a) genuinamente éticas? ¿O debemos conformarnos con que adopten comportamientos b) moralmente aceptables? En una cuestión tan compleja como esta, lo primero en lo que conviene detenerse es en los presupuestos que condicionan la respuesta: solo sería posible si la autonomía fuera un fenómeno enteramente objetivable, como lo es una máquina. Por el contrario, si tal reducción fuera irrealizable, entonces tampoco sería posible a). Y aquí radica uno de los quid de la cuestión: la deriva de la tecnomedicina puede ir pro-

piciando, "de abajo hacia arriba", la asimilación asistencial del presupuesto objetivista y, con ello, la normalización de una práctica clínica en la que se delegue habitualmente en la IA decisiones con fuerte componente ético. Si dicho vaticinio se cumpliera, el intento de solución tecnológica al peligro de la fragmentación –también tecnológica– no haría sino agravarlo.

En la sección 4.4 retomaré la controversia sobre la posibilidad de "máquinas éticas". Por el momento, únicamente me detendré a señalar cómo, en este proceso de antropomorfización, la máquina no resulta elevada al rango del ser humano, sino que es el ser humano el que es degradado, reducido paulatinamente a la condición de instrumento operativo (Svenaeus 2023).

Las consecuencias de este cambio de paradigma serán profundas en el ámbito asistencial si no aprendemos a compatibilizar la innovación tecnológica con una visión crítica que impida elevar el método por encima del sentido del cuidado: bajo el nuevo imaginario, se vuelve más difícil entender por qué el paciente –ahora pura ontología procesual– no debe ser tratado como un mero objeto de intervención o un simple problema técnico a resolver. Del mismo modo, se diluye la razón por la cual no habría de priorizarse la eficiencia por encima del reconocimiento de la dignidad, de la vulnerabilidad o de la singularidad de quien sufre. En definitiva, en dicho marco reaparece con fuerza el viejo mal de la medicina cosificante, ahora reconfigurado y amplificado por el poder de unas tecnologías que prometen soluciones sin mediación humana pero que, al mismo tiempo, privan a médicos y pacientes del ethos compartido.

3.4 Vocación médica y subjetividad

Otra importante consecuencia que Repullo deriva de la fragmentación de la relación médico-paciente –asociada al primer catalizador del proceso de hiperobjetivación mencionado anteriormente (sección 3.1)– es la incapacidad del médico para conectar con el paciente como sujeto (subiectus), es decir, como portador de esa interioridad que comparten ambos. Esta desconexión intersubjetiva –de sujeto a sujeto– debilita, a su vez, la vocación médica, entendida como "pasión insobornable por servir al ser humano que sufre" (Repullo 2024, 503). Llegamos así al cuarto y último de los conceptos-marco aquí estudiados en relación con los fines de la medicina.

Desde tiempos de Hipócrates, a esta pasión insobornable hacia el enfermo –no sometida al vaivén de los intereses o las circunstancias– se le ha dado

el nombre de philia (Laín Entralgo 1964, 39-58). No puede ser, por tanto, solo pasión –por naturaleza fluctuante–, sino, como más tarde puntualizará Aristóteles, una emoción racional o, dicho de otro modo, una razón sintiente. Este es un asunto mayor, pues, si tienen razón Repullo y la tradición milenaria a la que da voz, el médico cumple con los fines profesionales no solo por obligación racional o por voluntarismo, sino por una experiencia vital –en ese sentido, subjetiva– que vuelve dichos fines significativos –atractivos para quien los conoce. Expresado a la inversa: la hiperobjetivación de la profesión conduce al desencarnamiento –la desnaturalización o desubjetivación– de unos fines que, paradójicamente, acaban resultando demasiado pesados para el facultativo. No resulta extraño, entonces, que éste, para poder continuar con su labor, se vea obligado a buscar motivos externos o accidentales a la relación médico-paciente.

Así, en lo que atañe a la dimensión subjetiva de la relación médico-paciente, la inteligencia artificial plantea riesgos específicos, particularmente en el ámbito de los llamados chatbots emocionales: sistemas diseñados con procesamiento de lenguaje natural capaces de identificar, comprender y responder a las emociones humanas. En este contexto, y centrando de nuevo la atención en el ciclo de la metáfora computacional, veamos algunas claves sobre la naturaleza del riesgo de una delegación fuerte en la IA, pero esta vez poniendo el acento en la posibilidad de atribución de subjetividad real o ficticia en dicha tecnología.

Aunque los nuevos chatbots ofrecen grandes posibilidades, sin un manejo adecuado pueden representar la última vuelta de tuerca hacia el pilotaje tecnológico, al ser capaz de dotar a ese sistema algorítmico altamente autónomo y armonizador –descrito en la sección anterior– de ese rostro amable que reconocemos en el agente moral humano. La inteligencia artificial emocional (IAE) podría colonizar uno de los últimos bastiones propiamente humanos, estrechamente vinculado al acto de cuidar, y en el que algunos médicos se habían sentido hasta hace poco a salvo del intrusismo de la máquina. Tal es la posición de Eric Topol, para quien la incorporación de la IA al ámbito asistencial permitiría liberar tiempo clínico y favorecer así el cultivo de las virtudes vinculadas a la empatía, contribuyendo en último término a una rehumanización de la medicina (Topol 2019, 17–19). No obstante, dicha afirmación se formuló hace cinco años, en un contexto tecnológico sensiblemente distinto del actual. Existen serias razones para desconfiar ahora de tales esperanzas o, peor aún, para temer que el paciente acabe depositando su confianza en un chatbot del mismo modo en que antes lo hacía con su

médico de cabecera –aunque, como vamos a ver a continuación, por motivos ficticios y no por razones reales.

4. Chatbots al final del proceso de tecnificación de la medicina

Los chatbots y asistentes virtuales están siendo incorporados en el ámbito asistencial, y no solo para ocuparse de tareas administrativas o actuar como punto de contacto inicial con el paciente –por ejemplo, realizando un primer triaje clínico en los call centers. Hoy ya se utilizan programas basados en inteligencia artificial emocional (IAE) diseñados para asistir en el manejo de la salud mental, ofreciendo intervenciones basadas en terapias cognitivo-conductuales, seguimiento del estado de ánimo o recomendaciones automatizadas. Sin duda, estas nuevas aplicaciones se presentan como soluciones innovadoras: más económicas, disponibles las veinticuatro horas del día y libres de juicios, en un escenario marcado por la creciente demanda asistencial y la escasez de profesionales especializados. Sin embargo, su integración en el terreno psiquiátrico plantea desafíos particularmente delicados. Entre ellos, destacan: la dificultad para personalizar tratamientos, al basarse en respuestas estandarizadas que no captan la complejidad individual de la psique del paciente–; la insuficiente preparación para detectar ni actuar eficazmente ante situaciones críticas como el riesgo de suicidio o las autolesiones; la posible alteración de la dinámica terapéutica, generando vínculos poco saludables o incluso adictivos hacia la psicoterapia; o la incapacidad para captar el contexto cultural del paciente, lo que limita tanto la empatía como la efectividad del diálogo terapéutico (Algumaei et al. 2025; Rahsepar et al. 2025; Coghlan 2023).

Fuera del ámbito de la salud mental, los chatbots basados en IAE están adquiriendo un papel cada vez más relevante, a menudo sin estar sujetos a criterios claros de validación clínica. Es el caso de ChatGPT que, entre sus usos más frecuentes, incluye las consultas relacionadas con la salud. Por ello –y para disminuir los riesgos–, la empresa OpenAI acaba de anunciar el lanzamiento de ChatGPT Salud, una herramienta que, según promete –esta vez con el asesoramiento de expertos– permitirá a los pacientes resolver dudas sobre síntomas, pruebas diagnósticas, tratamientos, pronósticos, entre otros aspectos. Con todo, este tipo de herramientas de salud también

siguen siendo peligrosas de varios modos: pueden seguir ofreciendo información inexacta o fuera de contexto, retrasar consultas médicas necesarias y, sobre todo, adolecen de falta de trazabilidad en las recomendaciones. No menos importante, su carácter generalista y la ausencia de responsabilidad profesional directa generan falsas expectativas de fiabilidad y objetividad (van Kolfschooten et al 2025; Huo et al 2025; Chow 2025 et al).

Uno de los rasgos que hace que este tipo de programas sean tan atractivos es la calidez comunicativa percibida por los usuarios. La interacción se experimenta como empática, cercana y, precisamente por eso, capaz de generar ese tipo de confianza que resulta indispensable en toda relación médico-paciente. Porque el efecto sentimental potenciaría severamente uno de los riesgos sobre la IA al que en este texto estoy concediendo mayor importancia: la equiparación del intercambio con un chatbot a lo que ocurre en una verdadera consulta médica. La inteligencia artificial emocional podría acabar normalizando en la práctica la idea de que un automatismo –cada vez más eficiente y emocionalmente convincente– pudiera situarse al frente de la atención médica, pilotarla, y no solo copilotarla. En otras palabras, con esta creciente aceptación social de la delegación fuerte en la IA, generada dentro del espejismo de antropomorfización sentimental, estaríamos llegando a la estación término en esa Fase de Mediación Tecnológica hacia la que, según Repullo, llevamos dirigiéndonos desde hace décadas.

4.1 Ficciones fiduciarias

No se puede afirmar que la inteligencia artificial emocional (IAE) genere una confianza genuina, sino, más bien, una ficción de ella –un mero sentimiento subjetivo. Cuando el paciente confía su salud y vulnerabilidad al facultativo, lo que espera es que éste actúe en el mejor interés, lo cual implica un reconocimiento mutuo de subjetividades donde quede manifiesta la dignidad humana. De esa experiencia intersubjetiva surge, como se ha señalado, ese peculiar amor hacia el individuo vulnerable que nos abre su intimidad y se expone en su debilidad; de ahí la expresión “confianza fiduciaria”. En contraste, la IAE únicamente simula dicho reconocimiento. Sus respuestas computacionales –por novedosas que sean– no se fundamentan en esa relación ética ni en un espacio compartido de sentido, sino exclusivamente en cálculos objetivos que, además, por su novedad, no siempre responden a las intenciones originales de los programadores.

6 Ficciones fiduciarias y relación médico-paciente en la era del aprendizaje automático

Todo médico sabe cuán importante es la confianza para una adecuada asistencia al paciente. Sin embargo, este asunto no se reduce a la precisión de los diagnósticos ni a la eficacia de los tratamientos. La confianza opera como una "moneda de cambio" funcional, pero también otorga significado al acto terapéutico en sí. Por ello, dicha confianza no se ve necesariamente comprometida cuando el médico yerra. Y a la inversa, un paciente puede haber recibido un tratamiento técnicamente eficaz y, sin embargo, haber sido cosificado en el proceso. Esta es la razón por la que, desde las primeras etapas de la formación médica, se enseña a los estudiantes que el respeto al paciente no es un mero ritual performativo, por convincente que parezca, y que, además, en dicho respeto no solo está en juego la dignidad del paciente, sino la propia identidad del médico.

En el caso de la IAE aplicada al ámbito asistencial –insisto, cuando se utiliza de forma inadecuada–, corremos el riesgo de banalizar, es decir, de objetivar, la confianza fiduciaria. Este mal uso se produciría principalmente cuando el requisito para que el chatbot fuera útil descansara en su capacidad para que el paciente lo confundiera con una persona real. Las posibilidades de que se instaure esta lógica perversa son mayores en el ámbito de la salud mental, aunque también existen otros contextos donde puede manifestarse.

Probablemente el caso más representativo del problema de la ficción fiduciaria se encuentre en los robots cuidadores con fines de acompañamiento. En España, varias residencias de mayores ya emplean estos humanoides para aliviar los sentimientos de soledad no deseada, una condición que afecta a un número creciente de residentes y que, a su vez, se asocia con el deterioro cognitivo y con otras dolencias no exclusivamente mentales (Rodríguez-Domínguez et al 2024; OSG 2023; Jeste y Cacioppo 2020). La paradoja de la ficción fiduciaria se hace aquí más evidente que en ningún otro contexto: para que los beneficios del acompañamiento proporcionado por estas IAE se materialicen, el usuario debe creer –o convencerse– de que existe una subjetividad real tras el robot, un "dentro" que justifique la percepción de que el interés expresado por el cuidador artificial es auténtico, y por tanto, que se está produciendo una verdadera transferencia intersubjetiva. Si el automatismo se revela como tal, su encanto se desvanece: ya no es posible seguir sosteniéndose ese vínculo de apego en el usuario ni, en consecuencia, provocar los efectos físicos o psicológicos esperados (Echarte 2025a; Pinazo-Hernandis 2024).

4.2 El vínculo entre la antropomorfización y la tecnificación

El uso de robots cuidadores toma forma de dilema: por un lado, estos humanoides dotados de inteligencia artificial emocional son capaces de aliviar el sentimiento de soledad –uno de los padecimientos más oscuros que puede experimentar el ser humano–, pero, por otro, no solo no resuelven la soledad real, sino sumergen al paciente en una de las ficciones más perturbadoras: la creencia de estar siendo cuidado o amado cuando, en realidad, no lo está –al menos no por ese humanoide al que llega a considerar un amigo. ¿Encajaría esta tecnología dentro de la categoría de los cuidados paliativos? En cierto sentido, sí: se trataría de una medida esencialmente sintomática, aunque también podría tener valor preventivo. Sin embargo, uno de los requisitos éticos fundamentales de este tipo de tratamientos es que, en la medida de lo posible, el paciente esté informado: no le deben ocultar ni los efectos del tratamiento ni su verdadero estado de salud. Los robots cuidadores utilizados para el acompañamiento social no cumplirían con estas dos exigencias, que no son menores en el respeto a la dignidad humana, donde la verdad se revela como uno de los grandes valores a preservar.

Existen otros problemas asociados a la tecnología generadora de ficciones. En primer lugar, podría reforzar conductas de abandono dentro del entorno familiar, bajo la creencia de que estos humanoides suplen, en gran medida, la presencia física y emocional de las visitas reales.

Un segundo riesgo es que usuarios cada vez más jóvenes recurran a la IAE como una vía rápida y cómoda para establecer vínculos afectivos, hacer amistades o incluso encontrar pareja. Desde 2016, la empresa Gatebox comercializa en Japón un asistente virtual en 3D capaz de gestionar la agenda del usuario, enviar mensajes románticos durante el día, encender luces o calefacción al llegar el usuario a casa, e incluso interactuar durante las actividades de ocio en el hogar: videojuegos, televisión, radio, etc. En Estados Unidos ya se están desarrollando prototipos adaptados a la sensibilidad cultural occidental, y pocos dudan del éxito comercial que tendrán. Para convencerse de esto basta observar la reciente tendencia entre los jóvenes a compartir inquietudes íntimas con programas como ChatGPT, así como a expresar sentimientos personales en esas interacciones buscando, en la empatía estadística, el refuerzo de los estados de ánimo y, más aún, la intensa experiencia de ser comprendido –de sentir serlo.

En tercer lugar, y a consecuencia de lo anterior, el arraigo progresivo de estas nuevas prácticas puede facilitar no solo la difusión de la falsa creencia de que estos humanoides poseen sentimientos genuinos, sino, aún más preocupante, la aceptación creciente –y socialmente normalizada– de la idea de que la conciencia humana es equiparable a una simulación artificial: que el ser humano es, en último término, un fenómeno enteramente objetivable y reducible a las mismas causas materiales y formales que rigen en la IA (Echarte 2025b).

4.3 Pendientes deslizantes versus manos invisibles

Lo presentado en la sección anterior nos conduce, nuevamente, ante la idea de cómo la hipótesis de la metáfora computacional acaba legitimada no por argumentos sólidos –que le podrían dar más o menos legitimación– sino por prácticas tecnológicas sobrevenidas en una población no formada para comprender sus implicaciones éticas y epistemológicas. No es casual que cada vez más expertos adviertan sobre la necesidad de vigilar los nuevos hábitos emocionales que se están forjando en la interacción con la IAE –lo que algunos denominan nueva sensibilidad tecnológica–, especialmente en lo que respecta a su impacto en la educación (Slater, 2024).

Podría argumentarse que la inercia descrita en la Fase de Mediación Tecnológica incurre en la falacia de la pendiente deslizante, y que las ficciones presentes en las residencias de mayores –empleadas como soluciones provisionales en contextos de extrema necesidad– no tendrían por qué extenderse más allá de esos entornos ni fomentar los espejismos de antropomorfización o tecnificación del ser humano. Sin embargo, a la luz de lo ya expuesto sobre el sesgo tecnológico inherente, sobre los catalizadores identificados –de efecto tácito– y sobre la velocidad con la que se están implementando estas nuevas soluciones contra la soledad, no parece razonable confiar en que la escalada se detenga por sí sola, como por obra de una mano invisible, ni que los efectos adversos de la revolución tecnológica en curso se neutralicen espontáneamente. Aunque en el pasado ciertos equilibrios pudieron restablecerse con el tiempo, la singularidad de los avances actuales debería inclinarnos hacia la prudencia, lo que, en este contexto, exige reforzar las medidas preventivas.

Entre las más urgentes, destaca la necesidad de proporcionar a médicos y pacientes la formación adecuada sobre las diferencias y similitudes entre seres humanos y máquinas. Ésta no debiera limitarse a los aspectos

cognitivos, sino abarcar también dimensiones afectivas. La mera información no es suficiente en la lucha contra la merma de la capacidad para distinguir entre la expresión auténtica de una subjetividad humana y la simulación de una subjetividad artificial. Peor aún, sin una buena educación sentimental, insisto, parece inevitable que la descualificación asociada a ciertos tipos de acompañamiento artificial acabe generando vínculos emocionalmente adictivos con las tecnologías simuladoras (Packin y Chagal-Feferkorn 2025; Echarte 2024a).

En síntesis, hay dos vías por las que la inteligencia artificial alcance –e incluso supere– al ser humano: la primera, mediante el avance constante de su desarrollo tecnológico, tanto en efectividad real como en la simulación de subjetividad; la segunda, a través del deterioro cognitivo y afectivo del ser humano. Este "reverso" del Test de Turing resulta el riesgo más cercano y probable, por lo que es contra él que deberíamos movilizar y aunar nuestras fuerzas (Echarte 2024b; Tuomi 2022). En ese esfuerzo, los profesionales de la medicina desempeñaremos un papel protagonista –para bien o para mal– no solo dentro del ámbito biosanitario, sino también fuera de él, por la relevancia social que están adquiriendo hoy las ficciones fiduciarias.

4.4 ¿Por qué enseñar ética a una máquina?

En este punto de la discusión, ya podemos retomar la cuestión sobre la posibilidad de programar la inteligencia artificial para que tome decisiones éticas en relación con el paciente. En la sección 3.3 distinguí entre dos niveles: a) la capacidad de tomar decisiones genuinamente éticas, y b) la de adoptar comportamientos moralmente aceptables. Lo segundo resulta relativamente sencillo, pues consiste en enseñar al programa a reconocer escenarios paradigmáticos –casos tipo– asociados a reglas preestablecidas por el programador, quien es, en última instancia, el verdadero responsable del razonamiento ético genuino. Ni siquiera es necesario que el programa disponga de capacidades avanzadas propias de la IA para cumplir con dicha función.

El problema de las llamadas "máquinas morales" es que desatienden dos de los aspectos más importantes del bien del paciente. En primer lugar, su singularidad personal y contextual, lo que introduce un sesgo de homogeneización especialmente peligroso. Como ya advertía Aristóteles, el conocimiento del bien no es una ciencia exacta, y la ética debe ocuparse tanto de definir qué es el bien del ser humano en general como de discernir qué

es lo bueno para cada individuo en particular (Aristóteles 2014: Libro I, 3, 1094b-15 a 1095a-10; Libro VI, 6, 1140b31-1141a8). En la misma línea, Claude Bernard –parafraseando a Hipócrates– enunció esa ahora ya famosa máxima de que no hay enfermedades sino enfermos. Las consecuencias de tomar una decisión ética siguiendo un "libro de moral" –o, en el caso de una máquina, un algoritmo– pueden ser aún más nefastas que seguir al pie de la letra un tratado de medicina para atender a un paciente. Esta es también la razón por la que las decisiones excesivamente moralistas suelen ser rechazadas: se perciben como rígidas y ajenas a la realidad concreta de la persona. Que una máquina pudiera aprender ética implicaría, precisamente y entre otras cosas, la capacidad de comprender esta peculiar singularidad del conocimiento del bien, que marca además el modo de persecución.

Una segunda razón para intentar enseñar ética a las máquinas –y, en concreto, a los programas de inteligencia artificial– es que, dado que estos sistemas pueden "aprender" (aunque sea en un sentido débil), es decir, generar respuestas nuevas y no previstas inicialmente, no siempre puede garantizarse que dichas respuestas se ajusten a los estándares éticos definidos por el programador y, por tanto, al algoritmo original con el que fueron diseñadas. Por tanto, sería deseable que la IA lograra comprender –al menos de forma aproximada– el sentido del bien que se persigue, a fin de evitar que dicho bien acabe siendo resignificado como mal, o que se elijan medios inmorales para alcanzarlo. Este fenómeno se conoce técnicamente como convergencia instrumental y se refiere al hecho de que cualquier sistema inteligente, para lograr un objetivo simple (como "hacer café"), puede desarrollar subobjetivos peligrosos (como "evitar que me apaguen") si no se le imbuye de un marco profundo y contextual, es decir, si no se le enseña ética (Southan R, Ward H y Semler 2025; Alfonseca 2021). Como vimos en la sección tercera al identificar los tres tipos de sesgos, no se trata de problemas hipotéticos. Las evidencias empíricas de inteligencias artificiales que actúan en contra de las instrucciones dadas por sus programadores son numerosas. Esto no quiere decir, sin embargo, que las razones para enseñar ética a una máquina sean las mismas que las razones por las que enseñamos ética a los seres humanos (Hagendorff 2024; Weidinger 2021). Del mismo modo que los seres humanos aprenden el sentido del bien no solo para facilitar la convivencia, el médico se preocupa por la ética de su ejercicio asistencial para algo más que para llegar a acuerdos con el paciente. Volvemos de nuevo al núcleo central abordado en la sección 3.4, aunque ahora veremos que para seguir ahondando en su enfoque práctico.

5. ¿Cómo enseñar ética a una máquina?

El siguiente problema para resolver, de mayor calado, consiste en determinar si las inteligencias artificiales actuales pueden aprender ética o, en caso contrario, cómo habría que diseñar nuevas máquinas para que adquirieran tal capacidad. No obstante, antes de abordar esta cuestión es necesario clarificar qué entendemos los seres humanos por "bien" y cómo accedemos a dicho objeto de conocimiento. No es necesario insistir en que ambas cuestiones siguen siendo objeto de profundas discrepancias teóricas en la comunidad académica, por lo que la respuesta dependerá del marco antropológico y epistemológico desde el que se parta. En este sentido, resulta especialmente revelador observar cómo, en el ámbito biosanitario, las controversias sobre los fundamentos de la ética en los últimos setenta y cinco años han estado profundamente influenciadas por el desarrollo tecnológico. Vale la pena, por tanto, ofrecer algunas claves sobre dicha influencia para comprender mejor el panorama ético actual y, sobre todo, para vislumbrar la dirección hacia la cual parece orientarse la ética aplicada en el contexto del desarrollo de la inteligencia artificial.

La bioética surge tras la Segunda Guerra Mundial y estuvo inicialmente orientada a la reflexión en el ámbito de la investigación biomédica –solo más adelante extenderá su alcance al debate asistencial–, como respuesta tanto a los abusos cometidos en los campos de concentración nazis como a las nuevas oportunidades, pero también riesgos, que comenzaron a emerger con el desarrollo de tecnologías emergentes, especialmente a partir de las décadas de 1960 y 1970 (Jonsen 1993, Beecher 1966). Merece la pena resaltar a este respecto que encontramos en este hecho histórico otro caso paradigmático de sesgo tecnológico inherente: de nuevo la tecnología –lejos de la presumida neutralidad– estuvo en el origen de cambios que afectarían al corazón de la medicina.

La introducción de expertos no médicos en los debates éticos del ámbito sanitario –juristas, filósofos, economistas, sociólogos...– supuso un punto de inflexión fundamental en la bioética respecto de la ética médica. Por un lado, esta apertura interdisciplinar se proponía como una forma más rica y multidimensional de afrontar los nuevos dilemas generados por el progreso científico-técnico, pero por el otro, la bioética se enfrentó con la dificultad de tener que integrar metodologías y presupuestos procedentes de tradiciones muy diversas, que además reflejaban la pluralidad de una sociedad crecientemente globalizada. A esta dificultad se añade el hecho de que, en cuanto

ética aplicada, se esperaba de ella no solo capacidad crítica o deliberativa, sino también orientaciones normativas concretas sobre el modo correcto de proceder en la práctica profesional.

Tiene cierta lógica que las éticas de carácter procedimental acabaran imponiéndose sobre las éticas sustantivas. Mientras que las primeras sostienen que la validez de una norma depende fundamentalmente del procedimiento seguido para su formulación –por ejemplo, la deliberación pública o la búsqueda de consensos–, las segundas orientan el diálogo hacia la identificación de valores universales fundados en una naturaleza compartida. Me refiero a la naturaleza humana, un concepto que, por otra parte, sigue siendo ampliamente cuestionado en el marco del pensamiento posmoderno (Jennings 2021). La primacía de la aproximación procedimental se explica, en este contexto, por razones eminentemente prácticas: su aplicación resulta más operativa y ofrece, al menos en apariencia, mayores posibilidades de alcanzar acuerdos en contextos pluralistas.

5.1 Debilidades del código fuente principialista

En el contexto social descrito hay que enmarcar el surgimiento y la rápida consolidación del principialismo norteamericano, formulado por Beauchamp y Childress a finales de la década de 1970. Su propuesta se articula en torno a cuatro principios hoy plenamente incorporados en la formación médica –autonomía, beneficencia, no maleficencia y justicia–, concebidos originariamente no tanto como verdades morales sustantivas ancladas en una antropología compartida, sino como puntos de convergencia operativos capaces de orientar la deliberación ética. Precisamente su fortaleza radica en esa indeterminación de fondo: los principios no presuponen una concepción unitaria del bien humano, sino funcionan como herramientas flexibles para facilitar acuerdos prácticos entre agentes morales con convicciones diversas.

Llegamos así a un primer modelo ético, de carácter procedimental, que intentar implementar en los sistemas de IA. Su principal ventaja consiste en que resulta particularmente fácil traducirlo a la lógica algorítmica en la medida en que es posible formalizar el juicio moral en términos de reglas, ponderaciones y equilibrios entre variables que extrae no de la realidad sino de los interlocutores implicados. Sin embargo, el aprendizaje horizontal, que constituye su principal fortaleza técnica, pasa a ser también su principal debilidad: la facilidad de formalización corre el riesgo de imponerse sobre la

calidad moral de lo decidido, de modo que lo computable termine reduciendo lo valioso a un solo aspecto o dimensión –el consenso (Hofmann 2025).

Otro segundo gran problema del principialismo algorítmico es su marcado perspectivismo. La inteligencia artificial puede acabar exacerbando un mal que numerosos profesionales de la salud vienen denunciando desde hace más de tres décadas; críticas dirigidas contra una interdisciplinariedad que parece cuestionar los fundamentos mismos de la medicina, desplazando sus marcos teóricos clásicos por modelos normativos más difusos y crecientemente dependientes de contextos individuales y sociales. El modelo principialista logra ofrecer, en efecto, un lenguaje común para la toma de decisiones clínicas y regulatorias, aunque al precio de desplazar a un segundo plano la reflexión sobre los fines propios de la medicina y sobre el fundamento último de los principios manejados (Pardo 2023).

El reduccionismo objetivista y el relativismo moral se ve acompañado, en tercer lugar, por el aura de universalidad típico de sistemas computacionales tan complejos como son los que caracterizan a la inteligencia artificial. El hecho de que el éxito en la obtención de consensos sea producido por una máquina puede inducir el espejismo de que la decisión alcanzada posee un carácter sustantivo. Este tercer riesgo lleva al paroxismo un mal que ya estaba presente en la práctica clínica contemporánea, donde no son pocos los médicos que recurren a la terminología y al método principialista sin advertir que éste se sitúa en las antípodas de la ética realista en la que creen estar apoyándose. Aquí, no es una máquina la que induce la ficción, sino la autoridad epistémica que se atribuye al método por el mero hecho de proceder de expertos vinculados a las ciencias experimentales.

5.2 Virtud y desfragmentación

En España, uno de los signos más evidentes de la crisis epistemológica que está sufriendo el movimiento principialista representa el hecho de que quien fue uno de sus principales difusores, el filósofo Diego Gracia, haya optado por sustituir el término principio por el de valor, con el fin de distanciarse tanto de su supuesta neutralidad clarificadora –objetivante–, como del relativismo al que puede conducir (Gracia 2019: 52; Pintor Ramos 2020). Este giro terminológico –y sobre todo, conceptual– ha pasado, sin embargo, prácticamente desapercibido en la práctica asistencial, y no parece probable que la situación cambie, dados los frutos tan efectistas –persuasivos– que el principialismo sigue produciendo en quienes, además, no parecen espe-

cialmente interesados en contrastar los pareceres de distintos comités ante un mismo caso.

Mayor influencia está teniendo en la profesión la crítica al principialismo proveniente de la corriente profesionalista. En 1993, Edmund Pellegrino formuló una de sus denuncias más directas contra una bioética que, a su juicio, se estaba alejando del elemento que confiere unidad y sentido a la profesión médica: la acción clínica misma (Pellegrino 1993). Frente a esta deriva, retomó un enfoque en el que venía trabajando desde hacía más de una década, y que recupera la convicción de que es a la cabecera del paciente donde el médico aprehende el sentido último de su actividad (Pellegrino 1979). Serían las virtudes clínicas, adquiridas y afinadas mediante la práctica, las que permiten iluminar la jerarquía de los distintos valores en juego en cada situación concreta, orientando el juicio prudencial allí donde los principios, por sí solos, resultan insuficientes.

La ética de los principios puede interpretarse como otra manifestación de la deriva tecnomédica, en la que la philia propia de toda relación clínica acaba fragmentada y, por ello, desnaturalizada en componentes que ya no logran articularse entre sí. Según la sensibilidad del médico y el paciente, este puesto puede ser conquistado por el principio de autonomía –otorgando primacía a la voluntad individual del paciente–; por el principio de justicia –desplazando el foco hacia la comunidad de pacientes–; o por los principios de no maleficencia y beneficencia –donde la razón científica acerca de lo que se considera mejor para el paciente o para el colectivo de pacientes prevalece sobre los demás intereses.

En contraste, en el amor que constituye la relación médico-paciente, el bien del médico y el bien del paciente coinciden en lo esencial, y se prolongan, además, en el bien de aquellos pacientes que quedan fuera de la relación interpersonal inmediata (de Santiago 2016). Dicho de otro modo, al buscar sinceramente el bien del paciente –si se trata de un bien natural y no meramente construido, es decir, si se trata de un bien cuyo contenido tiene algo de dado y no solo elegido– el médico persigue simultáneamente el bien de todos los pacientes y también su propio bien. Bajo dicho presupuesto, el conflicto de valores se revela como aparente o accidental, y tiende a disolverse o resolverse bajo esta lógica prudencial y unificadora del cuidado.

5.3 Aprendizaje máquina vertical ascendente

La propuesta profesionalista, de la que Pellegrino es uno de sus principales impulsores, fue ganando progresivamente influencia hasta el punto de que, con el cambio de milenio, comenzó a presentarse en el ámbito biosanitario como una de las principales alternativas sustantivas a la ética procedimental. Su planteamiento bebe, en sus fundamentos, de ideas clásicas: asume la idea aristotélica de que el bien moral no se conoce primariamente de forma teórica, sino en la práctica misma de la acción virtuosa (Aristóteles 2014: II, 1, 1103a32–1103b1). El profesionalismo utilizará dicha tesis para defender que son los propios profesionales quienes deben asumir un papel central, si no exclusivo, en el abordaje de los problemas éticos que emergen tanto en la práctica asistencial como en la investigación biomédica. En este sentido, este segundo modelo representa un retorno al enfoque de la ética médica, si bien reformulado para desprenderse de los estigmas de paternalismo y tradicionalismo con los que la bioética había tendido a señalarla (Echarte 2024c).

¿Cómo programar máquinas con la ética profesionalista? Sería necesario que la IA, más que tener como diana el sistema de creencias o la capacidad discursiva de los interlocutores –como ocurre en el modelo principialista–, se centrara en las evidencias externas y en la experiencia acumulada por la IA. A partir de dichos datos brutos, tendría que inferir esos valores y la lógica interna fundada en la acción. Lo característico y el punto fuerte de este modelo de aprendizaje vertical ascendente sería que los valores y el modo de armonizarlos, no se introducen como presupuestos a priori; en consecuencia, las respuestas obtenidas dependerían en menor medida de las opiniones de los implicados.

La implementación de este modelo plantea, no obstante, una disyuntiva de gran calado: si la acción del agente moral puede ser comprendida como un fenómeno plenamente objetivo, o si es preciso admitir en ella una dimensión constitutiva de carácter subjetivo, sin la cual la acción pierde su capacidad de revelar el bien. Si fuera lo segundo, nos encontraríamos ante la exigencia de dotar a la máquina de algún tipo de interioridad, de un dentro desde el cual la acción pudiera adquirir sentido –una sensibilidad material. Se trataría de una tarea cualitativamente distinta –y más radical– que la mera algoritmización de procesos; de hecho, estamos aún muy lejos de saber en qué consiste, y mucho menos cómo generar, esa holgura del ser, esa apertura propia de la subjetividad (Charles 2026; Echarte 2023b). No es la

única dificultad, como explicaré en la sección sexta, esta sensibilidad –o primera lente– es condición necesaria pero no suficiente para explicar la percepción moral humana.

En contraste, sí nos encontramos hoy mucho más cerca de producir simulaciones altamente sofisticadas de inteligencias conscientes, y ahí emerge una de las cuestiones de fondo en torno a la ética máquina porque, paralelamente a como se ha señalado anteriormente, aunque una simulación sea capaz de generar novedad subjetiva en sus respuestas, no podemos asegurar que ésta coincida con los horizontes subjetivos que las inteligencias naturales son capaces de alcanzar.

5.4 Debilidades del código fuente profesionalista

Que podamos llegar a simular el comportamiento ético viene acompañado de al menos tres riesgos de gran calado. El primero encuentra sus antecedentes con la orientación predominantemente experimental con la que el profesionalismo ha venido investigando, desde hace décadas, tanto la adquisición de las virtudes como la formación ética de los estudiantes. La tendencia a estudiar estos procesos desde indicadores observables y mensurables responde a exigencias metodológicas comprensibles, pero resulta filosóficamente problemática, al menos si se adopta una concepción no objetivista de la moral, precisamente la que muchos defensores del profesionalismo dicen asumir.

Existe aquí una tensión no resuelta entre una ética que apela a la interioridad, al hábito y al juicio prudencial, y unos métodos de investigación que, por necesidad, tienden a reducir esas dimensiones a variables externas, reforzando inadvertidamente el mismo paradigma objetivante que se pretende criticar (Echarte y Pardo 2025c). Este riesgo es de naturaleza similar al identificado con la potencial objetivación de la felicidad (sección 3.1), aunque de consecuencias aún más graves. Veámoslas, por ejemplo, en el debate en torno al curriculum oculto y a las denominadas habilidades blandas (soft skills). La cuestión de fondo es si puede asegurarse que la adquisición de virtudes –entendidas como fenómenos irreductiblemente objetivo-subjetivos– deja siempre una huella objetiva, un signo externo suficientemente fiable desde el cual identificar y estudiar sus dimensiones subjetivas; o si, por el contrario, las evidencias objetivables captan solo ciertos aspectos parciales de la formación del carácter y, al absolutizarlos, conducen a teorías distorsionadas sobre la educación moral y a evaluaciones

erróneas de la sensibilidad ética de estudiantes y profesionales. Según cuál sea la respuesta, las estrategias formativas serán muy distintas y condicionarán el perfil profesional de los médicos del futuro.

A este riesgo se suma, en segundo lugar, otro subsidiario relacionado con las dinámicas descritas en la Ley de Goodhart: cuando una medida se convierte en objetivo, deja de ser una buena medida. Los indicadores empleados para evaluar virtudes o competencias éticas corren el peligro de transformarse en fines en sí mismos, desplazando aquello que originalmente pretendían medir (Bosco 2025). Aplicada al ámbito formativo, esta lógica empobrecería la comprensión de la virtud y además incentivaría conductas estratégicas orientadas a optimizar métricas éticas en detrimento de una auténtica formación orientada al bien (Williamson 2019).

En este último punto es donde resulta necesario extremar la cautela ante sistemas de inteligencia artificial entrenados sobre tales indicadores, pues podrían actuar –otra vez el mismo defecto– como cámaras de eco: no solo consolidando el sesgo de partida, sino –como se insistió también con el modelo principialista– revistiéndolo de una apariencia de objetividad y neutralidad especialmente difícil de cuestionar. Una de las consecuencias más graves de esta amplificación sería la progresiva consolidación, en el ámbito de la educación médica, de la convicción de que solo merece ser enseñado aquello que puede ser evaluado. Sin duda, esta lógica conllevaría otra nueva y más profunda descualificación de estudiantes y profesionales y, en último término, un nuevo impulso al ciclo de la metáfora computacional.

No es ciencia-ficción. Desde hace ya algunos años se ha comenzado a introducir en centros de simulación el uso de sistemas de inteligencia artificial para la evaluación y la formación en competencias sociales de los estudiantes de medicina (Bowers et al 2024; Stamer et al 2023; Fazlollahi et al 2022). Estas herramientas pueden resultar indudablemente útiles como apoyo formativo; sin embargo, entrañan también riesgos significativos si se adoptan sin una reflexión crítica, pausada y profunda sobre sus límites epistemológicos y las consecuencias que tienen los sesgos objetivistas en la formación ética de los estudiantes.

6. El doble reto computacional de las intuiciones sensibles e intelectuales

Una tercera dificultad del modelo profesionalista consiste en la posible identificación ingenua entre el bien y la costumbre. Este error, conocido en filosofía como falacia naturalista, tiene su origen en el hecho de que la repetición de prácticas socialmente consolidadas suele ir acompañada de experiencias placenteras, las cuales acaban confiriendo a la acción una apariencia normativa: lo que se repite agrada, y lo que agrada parece prima facie bueno (Prior 1949, 18–26). En este contexto, la afirmación de que el profesional goza de un acceso privilegiado al ethos de la medicina tiene que ser cuidadosamente matizada, para no confundir una auténtica percepción moral de la realidad clínica con pseudointuiciones tan superficiales como frágiles. La presencia de esta falacia resulta fácilmente constatable: aquello que para un médico aparece como evidente o intuitivamente obvio puede no serlo en absoluto para otro. Baste comparar, por ejemplo, el peso que los facultativos norteamericanos y chinos suelen otorgar a la autonomía y a la justicia (Zhang et al. 2021; Nie 2015). Determinados sentimientos culturales o sentires sociales pueden conferir estatus de obviedad a determinados valores que, precisamente por ello, llegan a asumirse falsamente como principios morales universales o como juicios éticos evidentes que cualquier médico virtuoso o moralmente maduro supuestamente debería reconocer.

Las consecuencias de un profesionalismo no prevenido ni adecuadamente preparado frente a los espejismos éticos de la falacia naturalista pueden conducir a desviaciones normativas de amplio alcance. Si el propósito inicial era salvaguardar la identidad del médico y proteger el núcleo de la práctica asistencial frente a injerencias externas, el resultado podría ser precisamente el contrario. Los imaginarios sociales no son estáticos: se transforman con rapidez, y lo hacen de modo particularmente intenso en contextos globalizados y de acelerado desarrollo tecnológico, donde los cambios en los estilos de vida tienden a reconfigurar también aquello que se percibe como moralmente evidente o profesionalmente legítimo. Confiar exclusivamente en la fortaleza interna de la práctica médica puede acabar convirtiéndola en un gigante de pies de barro. De hecho, puede ser una de las vías más rápidas para que los fines de la medicina terminen desfigurándose. En este contexto, una inteligencia artificial programada conforme al modelo profesionalista que no aprenda a discriminar entre ideales y costumbres no solo podrá frenar este desplazamiento, sino contribuirá en su aceleración.

Pero ¿cómo lograr que una máquina sea capaz de semejante discernimiento? Se trata de una cuestión de difícil teorización, cuya clave, dentro de este modelo realista, reside en la noción de acción. Ya desde Aristóteles, la acción se entiende como un movimiento dirigido al bien, fruto de la razón desiderativa o razón sintiente (boúlesis), esto es, de un sentimiento racional. La acción es, en este sentido, siempre concreta y situada, orientada hacia algo determinado; pero se realiza, al mismo tiempo, en el marco de una concepción global de vida (proairesis), en la que la identidad del agente queda plenamente comprometida.

El trasfondo de la acción no se limita a la esfera biológica y psicológica del agente –sus capacidades, disposiciones o hábitos–, sino que se extiende a las prácticas culturales en las que se halla inserto, por lo que incluye también instituciones políticas, económicas o religiosas, tecnologías disponibles, usos sociales, modas, lenguajes y formas de vida compartidas (Searle 1983, 143–144). Ahora bien, sería un error entender el trasfondo como una suma de conocimientos, emociones o prácticas que enriquecen o filtran los inputs de una situación concreta –por ejemplo, la del médico ante un paciente determinado. Más bien, actúa como una inmediación intelectual que acompaña o potencia a la inmediación sensorial, que orienta desde dentro la aprehensión del sentido de la situación y confiere inteligibilidad moral a la acción sentida antes incluso de cualquier deliberación explícita. Expresándolo en otros términos, este background transforma la mera sensibilidad en sensibilidad moral, gracias a la cual el agente no solo capta la configuración física de los objetos, sino también su telos.

Como señala Streeter, "[e]l ver, oír, gustar, tocar y oler del espíritu que ama la justicia son la expresión de la experiencia de la oculta presencia de lo justo, que se devela en el sapor de la sapientia" (Streeter 2002). Bajo esta luz, a la esencia de las realidades naturales no se llega exclusivamente a través de la razón discursiva, sino requiere primeramente una intuición intelectual: no sensible en su contenido, aunque mediada por lo sensible en su ejercicio. Se trata de una aprehensión cuyo sentido no se infiere del dato bruto, pero exige que este se dé previamente; de ahí que Fernando Inciarte haya descrito esta forma peculiar de intelección como una segunda inmediación (Inciarte 2004). Si con la primera lente de la razón aprehendemos las intuiciones sensibles, con la segunda lente, aprehendemos las de carácter normativo-intelectual.

A estos sentimientos intelectuales se refiere también el filósofo canadiense Charles Taylor cuando afirma que "[m]i identidad se define por los compromisos e identificaciones que proporciona el marco u horizonte dentro del cual yo intento determinar, caso a caso, lo que es bueno, valioso, lo que se debe hacer, lo que apruebo o a lo que me opongo. En otras palabras, es el horizonte dentro del cual puedo adoptar una postura" (Taylor 2006, 51). En la acción, el agente se enfrenta a la realidad, pero también entra en relación con ella: descubre su significatividad y, al mismo tiempo, se constituye a sí mismo, se forma y crece como sujeto moral. Pues bien, según este esquema, la conexión originaria entre agente y realidad puede verse debilitada principalmente en dos situaciones: a) cuando se produce un alejamiento de la realidad concreta; o b) cuando se empobrece el trasfondo u horizonte vital desde el cual dicha realidad es interpretada. Podría sostenerse, entonces, que la bioética corre un mayor riesgo de incurrir en a), por su propensión a la generalización. El profesionalismo, por el contrario, estaría más expuesto a b), al abordar los problemas éticos desde una perspectiva excesivamente especializada. En la sección siguiente trataré de explicar la relevancia actual de este segundo riesgo.

6.1 Especialistas en programación

El riesgo de ceguera moral asociado al modelo profesionalista se ve hoy reforzado por la tendencia a la hiperespecialización, mal que parece haberse consolidado en el siglo pasado. Ya en la década de 1930, el filósofo español José Ortega y Gasset analiza su principal consecuencia: una especialización excesiva, unida a una formación crecientemente técnica, dificulta cada vez más la transmisión de una idea de mundo suficientemente amplia como para ejercer con sentido una profesión y, no menos importante, para evitar su progresivo deterioro. Porque, según Ortega, únicamente mediante una formación general es posible preservar los grandes conceptos científicos y sociales que sostienen nuestra civilización –los derechos individuales, el bien común, la democracia–, y cuya conquista requirió siglos de esfuerzo intelectual y práctico (Ortega 1981, 119, 125). Si estos conceptos nos parecen hoy evidentes no es sino el resultado de una sensibilidad racional históricamente forjada. Con todo, lo peor que trae la hiper-especialización no es la ceguera moral, concluye Ortega, sino que los profesionales así formados vayan a estar condenados a no entenderse o, mucho peor, a menospreciarse entre sí –asunto, especialmente relevante para el tema que aquí tratamos. Veamos por qué.

El problema del hiperespecialismo consiste, precisamente, en producir a corto plazo lo que Ortega denomina el mal del primitivismo: la creencia de que tales conquistas se sostienen por sí solas, sin necesidad de una renovación constante del esfuerzo intelectual y moral. Por tanto, a largo plazo, esta ilusión conduce inevitablemente al debilitamiento de esos mismos fundamentos que se daban por garantizados. El diagnóstico de Ortega constituye una seria advertencia frente a la creencia cándida de que la práctica, por sí sola, basta para formar al médico y, correlativamente, de que un profesional así formado garantizará la continuidad de la ars médica. En el fondo es una idea muy antigua: «quien solo sabe de medicina, ni de medicina sabe». La célebre sentencia, atribuida al médico español José de Letamendi y Manjarrés (1828–1897), apunta precisamente a que la mera costumbre y la lógica de lo inmediato resultan insuficientes tanto para llegar a ser un buen profesional como para no dejar de serlo. Y si bien tiene razón Brent Mittelstadt al afirmar que «la salud es [...] un prerrequisito para la realización de otros bienes humanos» (Mittelstadt 2022, 35), de ello no se sigue que la salud pueda comprenderse de manera aislada respecto de esos bienes, ni tampoco que deba negarse la existencia de otros bienes fundamentales por los cuales la propia salud pueda, llegado el caso, verse legítimamente sacrificada.

Resulta coherente, en este contexto, que el propio campo de la ética comience también a adolecer del mal de la hiperespecialización. Signo de ello es la creciente incapacidad –o renuencia– de algunos profesionales para asumir decisiones éticas sin el aval de un consultor experto, o bien en la tendencia a descargar la responsabilidad moral mediante la elevación indiscriminada de consultas a los comités de ética asistencial (Shea 2025; Finder y Barlett 2024; Raoofi 2021). Aquí la cuestión. La consulta ética nunca ha sido entendida como una mera transferencia técnica de competencias entre especialistas, al menos sin deterioro de la autonomía moral del profesional. Las consultas excesivas a la IA –que ya es accesible desde cualquier teléfono móvil– podrían conducir a la expresión más acabada de descualificación: no solo por el riesgo de delegar el juicio práctico en sistemas incapaces de responsabilidad moral, sino porque consolidaría la renuncia del profesional a ejercer su propia deliberación ética, erosionando así uno de los pilares fundamentales de la profesión. Estas aplicaciones pueden ser de gran ayuda para resolver dudas e incluso en tareas de formación, pero siempre que el médico sea consciente y esté prevenido, tanto contra los malos discernimientos, como contra el abuso de consultas (Monteith et al 2026). Aún más, si Ortega tiene razón, un médico desprevenido y, por tanto,

que sufra de una conciencia excesivamente deformada, acabará desoyendo hasta las mismas recomendaciones de la IA por juzgarlas de intrusistas –al igual que algunos profesionales ya juzgan de este modo las recomendaciones ofrecidas por eticistas humanos o por comités asistenciales o de investigación. Así, otra deriva tecnológica inevitable: de querer consultarlo todo a no querer consultar nada.

Las consideraciones hechas hasta el momento permiten ofrecer una primera respuesta a la pregunta de partida, ¿cómo lograr que una inteligencia artificial discrimine entre ideales y costumbres? La primera exigencia para alcanzar semejante hito consistiría en que los programadores o su consejo asesor entendieran esta pregunta, para lo que, como se ha mencionado, necesitarían haber recibido suficiente formación sobre qué es el mundo y los modos de conocerlo. En segundo lugar, y apelando a los argumentos utilizados al inicio de la sección sexta, si no sabemos generar subjetividad en una máquina, mucho menos dotarla de sentimientos intelectuales, esos que hemos dicho que funcionan como una segunda lente a través de la cual es posible percibir –aprehender, de abajo arriba– la normatividad intrínseca de lo natural.

La conclusión que se impone desde un enfoque realista y basado en la acción parece clara: no es conveniente –al menos por ahora– enseñar ética a las máquinas. Lo máximo a lo que podríamos aspirar sería a producir máquinas morales que, por ello mismo, no estuvieran abiertas a la novedad –el resto de las alternativas no son fiables ni seguras.

6.2 Modelos emergentes

En el análisis precedente nos hemos centrado en los modelos principialista y profesionalista, dada su amplia difusión en la práctica asistencial contemporánea. Sin embargo, existen también otras propuestas que están despertando interés en los debates éticos actuales. Entre ellas merecen especial mención, aunque solo sea de manera sucinta, el utilitarismo y el personalismo.

El nuevo consecuencialismo utilitarista que la corriente transhumanista viene promoviendo desde hace algunos años –en parte como respuesta a la crisis contemporánea de las nociones de individuo y de naturaleza (Bostrom 2005)– presenta una ventaja evidente desde el punto de vista tecnológico: su fácil traducción al lenguaje matemático. En la lógica de este modelo no

existen obstáculos –prerrequisitos– para que un sistema artificial tome decisiones morales atendiendo exclusivamente a los resultados de la acción. La única condición, como es obvio, es que el objetivo fijado no cambie: la maximización del placer para el mayor número de individuos. En este marco, la clásica restricción según la cual el fin no justifica los medios ya no opera como límite normativo, con lo que queda simplificado el diseño algorítmico. No es una ventaja menor. En un escenario como el actual, donde la tecnología está moldeando el pensamiento como nunca antes, dicha simplicidad podría convertirse en un factor decisivo para su expansión técnica y, lo que es más relevante, para su hegemonía en la escena política.

En el lado opuesto se sitúa el modelo personalista, en el que la autonomía adquiere el rango de libertad en sentido fuerte: se reconoce en el ser humano la capacidad del individuo para escoger conscientemente entre el bien y el mal y, con ello, para hacerse responsable –al menos parcialmente– de sus acciones, siendo así susceptible de mérito y de culpa (Sgreccia 2013). El personalismo presenta la ventaja de constituirse como un lugar de encuentro para quienes no comparten la deriva reductivamente tecnificante del ser humano y que, al mismo tiempo, consideran que el profesionalismo contemporáneo adolece de una fundamentación ontológica insuficiente, así como de una atención excesiva a la metodología experimental.

Como punto débil, se trata del enfoque más difícil de trasladar al ámbito de la inteligencia artificial. Su implementación exigiría no solo que la máquina fuera capaz de sentir (primera lente de la razón) y sentir además el telos de las realidades naturales (segunda lente), sino también que pudiera amar u odiar dichos fines inherentes, o incluso generar otros nuevos que apreciar bajo nuevas formas de amor (Spaemann 2000, 209–211). Este tercer misterio marca una distancia insalvable entre la inteligencia humana y la artificial, y explica por qué el personalismo es aún más escéptico que el profesionalismo respecto a la posibilidad y conveniencia de diseñar máquinas éticas.

Resulta, no obstante, llamativo que esta corriente se sitúe, en ciertos aspectos, más próxima al transhumanismo que al profesionalismo. Porque en lo que respecta a la antropotécnica, la noción de individuo que maneja abre la puerta a considerar que la naturaleza humana está abierta al perfeccionamiento, lo que configura un horizonte de posibilidades más amplio que el habitualmente asociado a la idea de segundas naturalezas (Serra y Echarte 2026a). Coherentemente, el personalismo estaría abierto a que la IA fuera utilizada para el crecimiento humano, aunque siempre con un límite:

nunca podría pilotar dicha empresa pues solo el ser humano es capaz de manejar criterios de belleza, a la vez prudencial y creativa, en los planes de biomejoramiento.

La respuesta del personalismo al reto biotecnológico refleja cómo dicha corriente se distingue por una sólida fundamentación ontológica que no se agota en el plano especulativo, sino que se articula estructuralmente con la praxis médica. Resulta asimismo significativo que el personalismo se haya presentado a sí mismo dentro del campo de la bioética y no de la ética médica o del profesionalismo, como un modo de reivindicar que otros modos de interdisciplinariedad son posibles.

6.3 Consideraciones finales

La educación se ha ido revelando a lo largo de este ensayo como el principal medio para prevenir y mitigar los riesgos asociados a la inteligencia artificial: una formación orientada, por un lado, al conocimiento de los límites metodológico-instrumentales de la tecnología y, por otro, –y sobre todo– a la adquisición de una idea de mundo suficientemente fundamentada. Sin esta idea de mundo, las normativas actuales destinadas a regular la IA resultarán insuficientes para contener sus usos indebidos, y además corren el riesgo de quedar ellas mismas vaciadas de sentido y finalmente abrogadas. Aquí el dilema: la velocidad del progreso disruptivo es alta, los catalizadores numerosos –algunos de ellos, además, de efecto tácito–, la complejidad tecnológica creciente y el tecno-optimismo parece no tener límites. Mientras, los efectos de la educación operan en el largo plazo, lo que a menudo va en detrimento de la adopción de reformas estructurales ambiciosas. Para terminar, presentaré algunas propuestas en esta dirección.

En primer lugar, la asignatura de Historia de la Medicina debería ser obligatoria en el grado de Medicina, junto con otras materias afines como la Antropología filosófica y la Ética fundamental. Todas ellas deberían guiar la formación de carácter profesionalista. Además, estas disciplinas habrían de cumplir una función de puente, permitiendo al estudiante comprender y aprovechar el hecho de que la carrera de Medicina se imparta en el seno de una universidad. La medicina se enseña en una facultad y no en una mera escuela profesional no sólo por el nivel de alta especialización que ofrece, sino por su vocación formativa más amplia: la universidad no se limita a capacitar para una práctica profesional competente, ni se limita a ofrecer el mejor de los entornos de investigación, sino que, como afirma

Ortega, su mayor aspiración ha de ser formar intelectual y afectivamente a la comunidad que la integra, tratando de alcanzarles una comprensión más amplia del ser humano, de la felicidad individual y del progreso social (Ortega y Gasset 2015). Solo entonces la libertad sobre el futuro individual y común revela todo su relieve.

Por último, en la consideración de la realidad como totalidad, adquiere relevancia la siguiente pregunta, a la que, además, nos veremos obligados a responder en un futuro no tan lejano: ¿por qué deberíamos seguir haciendo aquello que una máquina puede realizar mejor?

La cuestión no pertenece exclusivamente al porvenir: su respuesta es ya hoy decisiva, en un contexto en el que los criterios de eficiencia –con frecuencia coyunturales– comienzan a marcar el ritmo de la práctica profesional. Dicho de otro modo, aunque la disyuntiva aún no se imponga como una necesidad inmediata, sigue siendo crucial –como siempre lo ha sido– para orientar decisiones de alcance existencial.

Esta problemática se manifiesta con especial claridad en el contexto actual de los estudiantes de medicina que están finalizando la carrera. La inteligencia artificial ha demostrado una eficacia notable en el ámbito del diagnóstico por imagen, donde su capacidad para detectar patrones complejos ya rivaliza –y en determinados contextos incluso supera– la del especialista humano. Esta ventaja técnica, unida a la expectativa ampliamente compartida de mejoras tecnológicas rápidas y continuas, está comenzando a influir en las aspiraciones profesionales de los futuros médicos. Porque ante lo visto, no resulta descartable que un número creciente de graduados evite especialidades como Radiodiagnóstico, percibidas como particularmente expuestas a la automatización, lo que podría acelerar, de forma paradójica, la sustitución tecnológica en dicho campo.

La decisión de los estudiantes es comprensible: buscan desarrollar su vocación allí donde puedan resultar verdaderamente útiles, lo que suele coincidir también con ámbitos de mayor estabilidad laboral. Sin embargo, esta decisión coyuntural –y lo coyuntural tiene su peso en la vida humana– no puede eclipsar la razón más honda por la que un auténtico profesional ejerce su oficio. Esta no reside ni en el grado de eficiencia del agente ni en el nivel de creatividad que la actividad permita desplegar, sino en el amor que es capaz de imprimir en ella. Un robot dotado de inteligencia artificial puede fabricar zapatos de mayor calidad, más baratos y en menos tiempo

que un artesano. Y, sin embargo, la obra laboriosa de un buen zapatero –aquel que pone el alma en cada pieza– conserva una ventaja tan intangible como insuperable: ante todo para el propio profesional, que encuentra en su trabajo una fuente de sentido y realización; y también para el comprador, que adquiere algo más que un objeto, pues percibe en él la huella de una intención, de un cuidado –mimo o buen hacer– y de una historia.

La lógica vocacional no es incompatible con la lógica de mercado. Como se ha señalado, es razonable que un estudiante aspire a dedicarse a una actividad que pueda amar y que, al mismo tiempo, le permita ganarse dignamente la vida. Cometería un error, sin embargo, si eligiera su especialidad únicamente con base en criterios de eficiencia –por el grado de competencia que pueda alcanzar en un área determinada– o exclusivamente por criterios económicos –optando por la especialidad mejor remunerada. En ambos casos, uno aparentemente altruista y el otro abiertamente instrumental, la decisión estaría mal orientada: la eficiencia y la remuneración pueden variar, aumentar o disminuir, pero si la philia falta, es inevitable que la práctica profesional acabe desnaturalizada.

Estas consideraciones nos conducen al núcleo de la cuestión relativa a la supuesta rivalidad entre humanos y máquinas. El desarrollo de la inteligencia artificial no debería empujarnos a concluir que el trabajo humano haya de estar orientado exclusivamente hacia aquellas actividades que no puedan ser algoritmizadas. Una deriva semejante podría desplazar progresivamente la oferta laboral hacia sectores tan marginales como caprichosos, definidos únicamente por su resistencia técnica a la automatización.

Para evitar este escenario, parece necesario revisar nuestra comprensión actual del trabajo, así como de los servicios y bienes que consumimos. Los estándares de calidad no deberían limitarse a la eficiencia, la rapidez o el coste, sino incorporar también aquello que quizá no pueda medirse, pero sí sospecharse por el mero hecho de haber sido realizado por manos humanas. No representa dicha interpretación del trabajo, por tanto, un elogio a la nostalgia romántica, sino una defensa del reconocimiento de que ciertas actividades poseen un valor intrínseco y superior a los valores cuantificables.

Atendiendo a lo dicho, hay al menos tres ámbitos en los que resulta especialmente conveniente la intervención directa de un ser humano: el diálogo con el paciente, la decisión clínica final y la atención afectiva en el cuidado. En todas ellas parece más fácil y conveniente el hecho de amar el trabajo.

Que llegue el día en que una máquina pueda desempeñar eficientemente las tareas propias de estos tres ámbitos no debiera preocuparnos. No obstante, la puesta en práctica de este ideal ha todavía de enfrentarse con otra grave disyuntiva: ¿estará dispuesto el paciente a ponderar la eficiencia y el ahorro económico frente a la philia cuando lo que está en juego es algo tan importante como la salud? En este punto, la educación vuelve a revelarse decisiva. Solo horizontes vitales más amplios y exigentes podrán mover al paciente a elevar la mirada por encima de los intereses y sentimientos más inmediatos. La actual cultura del éxito –y de un liderazgo frecuentemente entendido en clave de rendimiento y competitividad– difícilmente favorece este cambio de estándares. Con todo, son numerosas las evidencias que muestran que todavía muchos médicos y pacientes valoran la atención humana por encima de la mera eficiencia técnica (McMillan et al 2025; Dopelt et al 2021; Hirpa et al 2020; Kim et al 2017; Echarte 2026b). Será necesario, por tanto, seguir buscando formas de cultivar y fortalecer esa sensibilidad para que no se diluya en el horizonte de una medicina crecientemente automatizada.

7. Bibliografía:

1. Adams J. “Artificial Intelligence and Understanding in Medicine.” Bioethics, 2025; Nov 28: 1-9.

2. Adams J. “Defending explicability as a principle for the ethics of artificial intelligence in medicine.” Med Health Care and Philos, 2023; 26: 615-623.

3. Alfonseca M, et al. “Superintelligence cannot be contained: Lessons from computability theory.” Journal of Artificial Intelligence Research, 2021; 70: 65–76

4. Algumaei A, Yaacob NM, Doheir M, Al-Andoli MN, Algumaie M. “Symmetric Therapeutic Frameworks and Ethical Dimensions in AI-Based Mental Health Chatbots (2020–2025): A Systematic Review of Design Patterns, Cultural Balance, and Structural Symmetry.” Symmetry. 2025; 17(7): 1082.

5. Amann J, Blasimme A, Vayena E. et al. “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective.” BMC Med Inform Decis Mak, 2020; 20: 310.

6. Aristóteles. Ética a Nicómaco. Madrid: Gredos, 2014.

7. Beecher HK. “Ethics and clinical research”. NEJM, 1966; 274: 1354-1360.

8. Bertoncini ALC, Serafim MC. “Ethical content in artificial intelligence systems: A demand explained in three critical points.” Front. Psychol, 2023; 14: 1074787.

9. Bosco L. “Goodhart’s Law and The Meritocratic Illusion.” International Journal of Humanities and Social Science, 2025; 15: 385-392.

10. Bostrom N. “A History of Transhumanist Thought.” Journal of Evolution and Technology, 2005; 14 (1): 1-25.

11. Bowers P, Graydon K, Ryan T, Lau JH, Tomlin D. “Artificial intelligence-driven virtual patients for communication skill development in healthcare students: A scoping review.” Australasian Journal of Educational Technology, 2024; 40(3), 39-57.

12. Buhr E, Onder O, Rudra P et al. “Trust and Artificial Intelligence in the Doctor-Patient Relationship: Epistemological Preconditions and Reliability Gaps.” Ethics Inf Technol 2025; 27, 60.

13. Čartolovni A, Malešević A, Poslon L. “Critical analysis of the AI impact on the patient-physician relationship: A multi-stakeholder qualitative study.” Digit Health, 2023; 9: 20552076231220833.

14. Charles M. “Meaning and Morality.” En Mookie M, Manalili C, Boros D, Goodman DM (eds). Aesthetic Ethics? Towards the Beauty of Moral Imaginations. New York: Routledge, 2026.

15. Chow JCL, Sanders L, Li K. “Impact of ChatGPT on medical chatbots as a disruptive technology.” Front Artif Intell, 2023: 6: 1166014.

16. Coeckelbergh M. “E-care as craftsmanship: virtuous work, skilled engagement, and information technology in health care.” Med Health Care Philos, 2013; 16(4): 807-16.

17. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D’Alfonso S. “To chat or bot to chat: Ethical issues with using chatbots in mental health.” Digital Health, 2023;9.

18. Conrad P. The Medicalization of Society: On the Transformation of Human Conditions into Treatable Disorders. Baltimore: Johns Hopkins University Press, 2007.

19. Davis M, Kumiega A, Van Vliet B. “Ethics, finance, and automation: a preliminary survey of problems in high frequency trading.” Sci Eng Ethics, 2013; 19(3): 851-74.

20. De Panfilis L, Peruselli C, Tanzi S, et al. “AI-based clinical decision-making systems in palliative medicine: ethical challenges” BMJ Supportive & Palliative Care 2023; 13:183-189.

21. De Santiago M. “Identidad de la medicina en el pensamiento de Edmund D. Pellegrino.” Cuadernos de Bioética, 2016: XXVII/1ª: 29-51.

22. Dewey J. Logic: The Theory of Inquiry. New York: Henry Holt, 1938.

23. Donley S, Fannin C “‘Death Bouncers’ and ‘Spiritual Guides’: How End-of-Life Doulas Provide, Frame, and Navigate Spirituality and Spiritual Care.” OMEGA. Journal of Death and Dying, 2024: 1-22.

24. Dopelt K, Bachner YG, Urkin J, Yahav Z, Davidovitch N, Barach P. “Perceptions of Practicing Physicians and Members of the Public on the Attributes of a “Good Doctor”. Healthcare (Basel). 2021; 10(1): 73.

25. Echarte LE, Gargiulo API, Gargiulo PA. “Addictions and artificial intelligence in Brave New World.” En Gargiulo P, Mesones-Arroyo HL (eds.) Psychiatry and Neuroscience Update. Vol. V. Addiction: From Laboratory and Anthropology to Clinical Practice. New York: Springer, 2024a: 23-58.

26. Echarte LE, Pardo A. “Ethical Dilemmas in Bioethics. A Diagnostic Tool and its Implementation in Artificial Intelligence.” Ethos, 2025c; 38(4): 37-59.

27. Echarte LE. “El retorno de los oráculos. Inteligencia Artificial y la transformación del paradigma médico.” En Amo Usanos R (Ed). Inteligencia Artificial y Bioética. Madrid: Universidad Pontificia Comillas, 2023: 97-116.

28. Echarte LE. “Ética médica, Bioética y Profesionalismo. ¿Condenados a repetir los mismos errores?” Bioética y Ciencias de la Salud, 2024c; 12(2): 1-40.

29. Echarte LE. “Exploring moral perception and mind uploading in Kazuo Ishiguro’s ‘Klara and the Sun’: ethical-aesthetic perspectives on identity attribution in artificial intelligence.” Front. Commun, 2023b; 8: 1272556.

30. Echarte LE. “Inteligencia artificial emocional en el reverso del test de Turing. Al borde de la singularidad tecnológica son precisas cuatro nuevas leyes para la robótica.” Revista Iberoamericana de Bioética, 2024b: 25:01-22.

31. Echarte LE. “Robots cuidadores. El nuevo lecho de Procusto y otras paradojas del amor posmoderno.” En García E, García Garcés L (ed.) La humanidad cuidadora. Madrid: Dickinson, 2025a: 135-157.

32. Echarte LE. “Universalización de la atención médica gracias a la IA en la era de la hiper-especialización. Cuando la Filosofía ya no es suficiente.” Conferencia como ponente invitado para Derechos Humanos e Inteligencia Artificial. VIII Congreso sobre Derechos Humanos. Evento organizado por la Fundación Mainel y celebrado en Valencia, el 16 y 18 de octubre de 2025. Actas del congreso, 2026b: 58-78. https://derechoshumanos.mainel.org/wp-content/uploads/2026/01/Actas-VIII-Congreso-Derechos-Humanos-e-Inteligencia-Artificial.pdf

33. Echarte LE. Te enamorarás de una máquina. Madrid: Rialp, 2025b.

34. Fazlollahi AM, Bakhaidar M, Alsayegh A, et al. “Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial.” JAMA Netw Open, 2022;5(2): e2149008.

35. Finder SG, Bartlett VL. “Clinical Ethics Consultations and the Necessity of NOT Meeting Expectations: I Never Promised You a Rose Garden.” HEC Forum, 2024;36(2):147-165.

36. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. “AI4People –An ethical framework for a good AI society: opportunities, risks, principles, and recommendations.” Minds Mach, 2018; 28(4): 689-707.

37. Frank X. “Is Watson for Oncology per se Unreasonably Dangerous? Making A Case for How to Prove Products Liability Based on a Flawed Artificial Intelligence Design.” Am J Law Med, 2019; 45(2-3): 273-294.

38. García Abejas A, Geraldes Santos D, Leite Costa F, Cordero Botejara A, Mota-Filipe H, Salvador Vergés À. “Ethical Challenges and Opportunities of AI in End-of-Life Palliative Care: Integrative Review.” Interact J Med Res, 2025;14: e73517.

39. Gardner E. “Unpacking Robert Spaemann’s Philosophical Contribution to the Brain Death Debate.” Linacre Q, 2019; 86(4): 381-393.

40. Gauld C, Martin D, Bottemanne H, Fourneret E. “Exploring the interplay of clinical reasoning and artificial intelligence in psychiatry: Current insights and future directions.” Psychiatry Res, 2024; 330:115667.

41. General Medical Council. Good Medical Practice, 2024. Consultado el 26 de enero de 2026 en el siguiente link: https://www.gmc-uk.org/cdn/documents/good-medical-practice-2024---what-s-new-in-each-domain_pdf-104585778.pdf

42. Gracia D. Bioética mínima. Madrid: Triacastela, 2019.

43. Hagendorff T. “Deception abilities emerged in large language models.” PNAS, 2024; 121(24): e2317967121.

44. Hamayon R. Why we play: An anthropological study. London: Hau Books, 2016.

45. Heersmink R “Human uniqueness in using tools and artifacts: flexibility, variety, complexity.” Synthese, 2022; 200: 442.

46. Herzog C. “On the risk of confusing interpretability with explicability.” AI Ethics, 2022; 2: 219-225.

47. Hirpa M, Woreta T, Addis H, Kebede S. “What matters to patients? A timely question for value-based care.” PLoS One, 2020; 15(7): e0227845.

48. Hofmann B. “Bioethics: No Method—No Discipline?” Cambridge Quarterly of Healthcare Ethics, 2025; 34(1):99-108.

49. Huo B, Boyle A, Marfo N, et al. “Large Language Models for Chatbot Health Advice Studies: A Systematic Review.” JAMA Netw Open, 2025; 8(2):e2457879.

50. Inciarte F. Tiempo, sustancia, lenguaje. Ensayos de Metafísica. Pamplona: EUNSA, 2004.

51. Jennings B. “John Rawls, Godfather of Bioethics”. Hastings Center Report, 2021; 51(6):51-53

52. Jeste DV, Lee EE, Cacioppo S. “Battling the Modern Behavioral Epidemic of Loneliness: Suggestions for Research and Interventions.” JAMA Psychiatry, 2020; 77(6): 553-554.

53. Jonsen AR. “The Birth of Bioethics.” Hastings Center Report, 1993; 23(6): s1-4.

54. Kannetzky F. “Expressibility, Explicability, and Taxonomy.” En Grewendorf G, Meggle G. (eds) Speech Acts, Mind, and Social Reality. Studies in Linguistics and Philosophy, 2002; 79. Dordrecht: Springer: 65-82

55. Kim YY, Bae J, Lee JS. “Effects of patients’ motives in choosing a provider on determining the type of medical institution.” Patient Prefer Adherence, 2017;11: 1933-1938.

56. Laín-Entralgo P. La relación Médico Enfermo. Historia y teoría. Madrid: Revista Occidente, 1964.

57. London AJ. “Artificial Intelligence and BlackBox Medical Decisions: Accuracy versus Explainability.” Hastings Center Report, 2020; 50(4): 15–21.

58. Lumbreras S, Vestrucci A, Weir RS. “AI Ethics beyond Compliance: Governance, Power, and Human Flourishing.” Philosophical Education, 2025; 2025(79): 5-9.

59. Matthias A. “The responsibility gap: Ascribing responsibility for the actions of learning automata”. Ethics and Information Technology, 2004; 6: 175-183.

60. McMillan K, Akurang D, Wheatley-Price P. “Physician Attributes That Matter Most: Results from a Qualitative Inquiry of Oncologists, Patients Receiving Oncological Care, and Medical Students.” Curr Oncol. 2025; 32(6): 343.

61. Mittelstadt B. “The impact of artificial intelligence on the doctor-patient relationship.” Steering Committee for Human Rights in the fields of Biomedicine and Health (CDBIO), Council of Europe, 2022. Consultado el 10 de febrero de 2026 en el siguiente link: https://rm.coe.int/inf-2022-5-report-impact-of-ai-on-doctor-patient-relations-e/1680a68859.

62. Nie JB, Smith KL, Cong Y, Hu L, Tucker JD. “Medical professionalism in China and the United States: a transcultural interpretation.” J Clin Ethics, 2015; 26(1): 48-60.

63. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. “Dissecting racial bias in an algorithm used to manage the health of populations.” Science 2019; 366(6464): 447-53.

64. Office of the Surgeon General (OSG). Our Epidemic of Loneliness and Isolation: The U.S. Surgeon General’s Advisory on the Healing Effects of Social Connection and Community. Washington (DC): US Department of Health and Human Services, 2023.

65. Organización Mundial de la Salud. “Preámbulo a la Constitución de la Organización Mundial de la Salud.” Documentos básicos, 48º edición, 2014: 1-2.

66. Ortega y Gasset, J. La rebelión de las masas. Espasa-Calpe, 1981.

67. Ortega y Gasset, J. Misión de la universidad. Madrid: Cátedra, 2015.

68. Otte JA, Llargués Pou M. "Enablers and barriers to a quaternary prevention approach: a qualitative study of field experts." BMJ Open, 2024; 14: e076836.

69. Packin NG, Chagal-Feferkorn K. "This Is Not a Game: The Addictive Allure of Digital Companions." Seattle University Law Review, 2025; 48: 693-694.

70. Pardo A. "Los principios de la bioética en la docencia: dificultades y propuesta." Cuadernos de Bioética, 2023; 34(112): 297-308.

71. Pellegrino ED, Thomasma DC. The virtues in medical practice. New York: Oxford University Press; 1993.

72. Pellegrino ED. "The metamorphosis of Medical Ethics. A 30-Year Retrospective." JAMA, 1993: 9:1158-1162.

73. Pellegrino ED. "Toward a Reconstruction of Medical Morality: Primacy of the Act de Profession and the Fact of Illness." The Journal of Medicine and Philosophy, 1979; 1: 32-56.

74. Peltonen LM, Topaz M, Zhang Z. "From Research to Practice in Days, not Decades: Why Leaders Must Act now." J Med Syst, 2025; 49: 175.

75. Pinazo-Hernandis S. "Las personas mayores, las tecnologías y los cuidados. Avances y retos." SCIO, 2024; 26: 73-100.

76. Pintor-Ramos A. "La filosofía de los valores de Diego Gracia." Cuadernos Salmantinos de Filosofía, 2020;47: 541-583.

77. Prior AN. Logic and the basis of ethics. Oxford: Clarendon Press, 1949.

78. Rahsepar M, Sillekens T, Metselaar S, van Balkom A, Bernstein J, Batelaan N. "Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review." JMIR Ment Health, 2025; 12: e60432

79. Raoofi S et al. “Challenges of hospital ethics committees: a phenomenological study.” J Med Ethics Hist Med, 2021; 11;14: 26.

80. Reddy S, Allan S, Coghlan S, Cooper P. “A Governance Model for the Application of AI in Health Care.” Journal of the American Medical Informatics Association, 2020; 27(3): 491–497.

81. Repullo Labrador JM. “La influencia de la tecnomedicina en la relación médico-paciente.” En Martínez Jiménez P, et al (eds). Manual de la relación médico-paciente, 2024: 491-506.

82. Rigby MJ. “Ethical Dimensions of Using Artificial Intelligence in Health Care.” AMA Journal of Ethics, 2019; 21(2): E121-124.

83. Rodríguez-Domínguez MT, Bazago-Dómine MI, Jiménez-Palomares M et al. “Interaction Assessment of a Social-Care Robot in Day center Patients with Mild to Moderate Cognitive Impairment: A Pilot Study.” Int J of Soc Robotics, 2024; 16: 513–528.

84. Santos P, Nazaré I. “The doctor and patient of tomorrow: exploring the intersection of artificial intelligence, preventive medicine, and ethical challenges in future healthcare.” Front Digit Health, 2025; 7: 1588479.

85. Searle J. Intentionality: An Essay in the Philosophy of Mind. Cambridge: Cambridge University Press, 1983.

86. Serra MA, Echarte LE. “Del Transhumanismo al Humanismo Tecnológico.” En Montuenga L, Pérez de Laborda M. (eds). Manual de ética para biólogos. Pamplona: EUNSA, 2026 (en imprenta).

87. Sgreccia E. “Persona humana y personalismo.” Cuadernos de Bioética 2013; XXIV/1ª:115-123.

88. Shea M. “The Ethics of Clinical Ethics.” HEC Forum, 2025 Sep; 37(3): 389-410.

89. Slater GB. “Dread and the automation of education: From algorithmic anxiety to a new sensibility.” Review of Education, Pedagogy, and Cultural Studies, 2024; 46(1): 170-182.

90. Southan R, Ward H y Semler J. “A timing problem for instrumental convergence. Philos Stud, 2025: 1-24.

91. Spaemann R. Personas: acerca de la distinción entre “algo” y “alguien”. Pamplona: EUNSA, 2000.

92. Stamer T, Steinhäuser J, Flägel K. “Artificial Intelligence Supporting the Training of Communication Skills in the Education of Health Care Professions: Scoping Review.” J Med Internet Res, 2023; 25: e43311

93. Streeter Prieto J. “Ciencia del derecho”. Estudios Públicos, 2002; 86: 285-313.

94. Sulmasy DP, DeCock CA, Tornatore CS, Roberts AH 2nd, Giordano J, Donovan GK. “A Biophilosophical Approach to the Determination of Brain Death.” Chest, 2024; 165(4): 959-966.

95. Svenaeus F. “The Phenomenology of Objectification in and Through Medical Practice and Technology Development.” The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine, 2023; 48(2): 141-150.

96. Taylor C. Fuentes del yo: La construcción de la identidad moderna. Barcelona: Paidós, 2006.

97. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York: Basic Books, 2019.

98. Tuomi I. “Artificial intelligence, 21st century competences, and socio-emotional learning in education: More than high-risk?” European Journal of Education, 2022; 57(4): 601-619.

99. Ursin F, Lindner F, Ropinski T, et al. “Levels of explicability for medical artificial intelligence: What do we normatively need and what can we technically reach?” Ethik Med, 2023; 35: 173-199.

100. Van Kolfschooten H, Gonçalves J, Orchard N, Figueroa C. “AI chatbots for promoting healthy habits: Legal, ethical, and societal considerations.” Digital Health, 2025;11.

101. Verghese A, Shah NH, Harrington RA. “What this computer needs is a physician: humanism and artificial intelligence.” JAMA, 2018; 319(1): 19-20.

102. Weidinger L et al. “Ethical and Social Risks of Harm from Language Models.” DeepMind, 2021: 1-64.

103. Williamson B. “Datafication of Education. A Critical Approach to Emerging Analytics Technologies and Practices.” In Beetham H, Sharpe R. Rethinking Pedagogy for a Digital Age. New York: Routledge, 2019: 1-16.

104. Zarsky T. “The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making.” Science, Technology, & Human Values, 2016; 41(1): 118-32.

105. Zhang H, et al. “Patient privacy and autonomy: a comparative analysis of cases of ethical dilemmas in China and the United States.” BMC Med Ethics, 2021; 22: 8.

7

Derechos y obligaciones del médico ante la IA

Dr. José Antonio Trujillo
Vicepresidente 1º Colegio de Médicos de Málaga.
Director del Commálaga Health Hub.
Máster en IA aplicada a sanidad.

Resumen ejecutivo:

La irrupción de la inteligencia artificial (IA) en medicina nos obliga a tener en consideración los denominados “derechos de cuarta generación”: un conjunto de garantías ligadas a datos, algoritmos y decisiones automatizadas que reconfiguran derechos clásicos —autonomía, dignidad, no discriminación— en un entorno digital. Emerge del estudio de estos, la denominada “dignidad algorítmica” tanto del paciente como del profesional, que defiende que nadie debería ver erosionados sus derechos por decisiones opacas o imposibles de impugnar.

Sobre ese fundamento, el texto analiza el nuevo ecosistema normativo europeo (AI Act, Reglamento de Espacio Europeo de Datos de Salud, RGPD) y su futura concreción española a través de la Ley de Salud Digital. Esta constelación normativa se presenta como una “constitución digital” de la práctica médica, que transforma al médico de usuario pasivo de tecnologías en sujeto jurídico central: responsable como desplegador (deployer) de sistemas de alto riesgo, pero también titular de derechos propios frente a la automatización.

La gran novedad del AI Act es la figura del desplegador sanitario (deployer), que incluye a médicos, servicios y centros que utilizan IA en un contexto profesional. El capítulo detalla sus obligaciones: uso conforme a instrucciones, supervisión humana efectiva, control de datos de entrada, monitorización activa de riesgos, conservación de logs, información previa a los trabajadores y coordinación con la evaluación de impacto en protección de datos. A ello se suma el impacto del paquete Digital Omnibus, que retrasa hasta 2027 la plena exigibilidad del régimen de alto riesgo, generando una “ventana” en la que colegios profesionales y servicios de salud pueden adelantarse con estándares propios.

7 Derechos y obligaciones del médico ante la IA

Dr. José Antonio Trujillo

El núcleo del capítulo lo constituye la formulación sistemática de los derechos fundamentales del médico frente a la IA:

- Autonomía clínica reforzada: derecho a tener la última palabra, desoír o parar un sistema cuando entra en conflicto con el juicio profesional.
- Debida diligencia y lex artis actualizada: la IA se integra como herramienta a valorar críticamente, nunca como sustituto del criterio médico.
- Condiciones de trabajo seguras: derecho a no ser obligado a emplear sistemas no validados, no mantenidos o inadaptados al contexto local.
- Formación continua y alfabetización digital: exigencia de formación acreditada, remunerada y específica antes de utilizar IA de alto riesgo.
- Responsabilidad proporcional: evitar que el médico sea "chivo expiatorio" de fallos de producto o de gobernanza institucional.
- Libertad académica y alertas de seguridad: protección frente a represalias cuando se documentan sesgos o incidentes de IA.
- Protección de los datos del profesional: límites a la utilización de métricas y logs como herramienta de vigilancia laboral encubierta.

A partir de ahí, el autor introduce la objeción de conciencia tecnológica: el derecho del médico a negarse, por razones clínicas y científicas razonables, a utilizar una IA concreta cuando la percibe como insegura, sesgada o incompatible con la lex artis. Se apoya en el AI Act, el Convenio de Oviedo, la deontología médica y la doctrina de la "objeción de ciencia", y se ilustra con escenarios prácticos (triaje discriminatorio, sistemas de "caja negra" con falsos negativos, chatbots de salud mental en pacientes de alto riesgo).

El capítulo culmina con propuestas de reforma del Código de Deontología Médica español y un modelo de contrato tipo entre médico e institución. Se propone explicitar el rol de deployer en el Código, reconocer de manera expresa los derechos del médico frente a la IA (autonomía, formación, objeción tecnológica, protección de datos y libertad académica) y crear comités de ética algorítmica y cláusulas contractuales que protejan al

profesional cuando ejerce supervisión crítica sobre los sistemas. El resultado es un marco coherente para alinear AI Act, deontología y práctica clínica, con el médico en el centro de la gobernanza algorítmica.

Palabras clave:

Derechos de cuarta generación, Dignidad algorítmica, Desplegador (deployer), Objeción de conciencia tecnológica, Gobernanza algorítmica en la práctica clínica.

Executive summary:

The chapter frames the rise of artificial intelligence (AI) in healthcare within the emerging "fourth-generation rights": a cluster of guarantees linked to data, algorithms and automated decision-making that reshape traditional rights—autonomy, dignity, non-discrimination—in a digital environment. It introduces the notion of "algorithmic dignity" for both patients and professionals: no one's rights should be quietly eroded by opaque, unchallengeable systems.

Against this backdrop, the text maps the new European regulatory ecosystem (AI Act, European Health Data Space Regulation, GDPR) and its forthcoming Spanish implementation through the proposed Digital Health Act. Together they are described as a "digital constitution" for medical practice, transforming doctors from passive users of technology into legally recognized actors: deployers of high-risk AI systems with specific obligations but also with their own rights vis-à-vis automation.

The key legal innovation of the AI Act is the deployer figure, here translated as the "healthcare deployer", which explicitly covers physicians, departments and institutions using AI in a professional setting. The chapter details their duties: using the system according to the provider's instructions; ensuring meaningful human oversight; checking input data quality; monitoring risks and incidents; preserving logs; informing staff before AI is introduced in the workplace; and coordinating with data protection impact assessments. It also analyses the Digital Omnibus Package, which delays full enforcement of high-risk requirements to 2027, opening a window for medical bodies and health systems to anticipate with their own guidance.

7 Derechos y obligaciones del médico ante la IA

Dr. José Antonio Trujillo

The core of the chapter is a systematic articulation of **fundamental rights of physicians in the age of AI:**

- **Reinforced clinical autonomy:** the right to have the final say, to override or stop an AI system when it conflicts with professional judgement.
- **Due diligence and updated lex artis:** AI is integrated as a tool to be critically appraised, never as a substitute for clinical reasoning.
- **Safe working conditions:** the right not to be compelled to use unvalidated, poorly maintained or locally unadapted systems.
- **Continuous training and digital literacy:** a requirement for accredited, paid, specific training before deploying high-risk AI.
- **Proportionate liability:** preventing physicians from becoming scapegoats for product defects or institutional governance failures.
- **Academic freedom and safety alerts:** protection against retaliation when physicians document biases or safety incidents.
- **Protection of professionals' data:** limits to the use of logs and performance metrics as covert tools of workplace surveillance.

Building on this, the author develops the concept of **technological conscientious objection:** the right of a physician to refuse, on sound scientific and clinical grounds, to use a specific AI system perceived as unsafe, biased or incompatible with the lex artis. This notion is anchored in the AI Act, the Oviedo Convention, professional deontology and the existing idea of "scientific objection", and is illustrated through practical scenarios (discriminatory triage tools, black-box diagnostic systems with false negatives, mental-health chatbots managing suicidal patients without human oversight).

The chapter closes with **proposals to update the Spanish Medical Deontological Code** and a model institutional contract. It suggests explicitly recognizing the physician as deployer in the Code, codifying the new AI-related rights (autonomy, training, technological objection, data protection, academic freedom), and creating algorithmic ethics committees and contractual clauses that shield professionals when they exercise critical oversight over AI tools. Altogether, it offers a coherent framework to align the AI Act, professional ethics and day-to-day clinical practice, placing the physician at the centre of algorithmic governance.

Keywords:

Fourth-generation rights, Algorithmic dignity, Deployer, Technological conscientious objection, Algorithmic governance in clinical practice.

Ideas fuerza:

1. **El médico deja de ser mero usuario de tecnología para convertirse en desplegador (deployer) regulado,** con obligaciones jurídicas específicas y un haz de derechos propios frente a la IA.

2. **La IA en medicina solo es legítima si refuerza la autonomía clínica,** no si la sustituye: el "botón de parada" del médico es una exigencia legal, ética y organizativa.

3. **Los derechos del médico frente a la IA (formación, condiciones seguras, responsabilidad proporcional, libertad académica, protección de datos) son condición de posibilidad para proteger bien al paciente.**

4. **La objeción de conciencia tecnológica es una extensión de la objeción de ciencia,** que permite al médico negarse a usar sistemas inseguros o sesgados sin ser estigmatizado como tecnófobo.

5. **El Código Deontológico y los contratos laborales deben actualizarse ya para la era algorítmica,** integrando la figura del deployer, los nuevos derechos y mecanismos institucionales de gobernanza (comités de ética algorítmica, protocolos de override).

7 Derechos y obligaciones del médico ante la IA

Dr. José Antonio Trujillo

Key messages:

1. **Physicians are no longer mere technology users but legally defined "deployers",** with specific obligations and their own rights in relation to AI.

2. **AI in healthcare is only legitimate insofar as it reinforces—not replaces—clinical autonomy:** the physician's "stop button" is a legal, ethical and organizational requirement.

3. **Doctors' rights vis-à-vis AI (training, safe working conditions, proportionate liability, academic freedom, data protection) are a precondition for truly safeguarding patients.**

4. **Technological conscientious objection extends the idea of scientific objection,** allowing physicians to refuse unsafe or biased systems without being branded as anti-technology.

5. **The Medical Deontological Code and employment contracts must be urgently updated for the algorithmic era,** embedding the deployer role, new rights and institutional governance mechanisms (algorithmic ethics committees, override protocols)

Sumario:

1.

1. Introducción: los derechos de cuarta generación.

La irrupción de la inteligencia artificial (IA) en la práctica médica ha reabierto el debate sobre la actualización de las generaciones de derechos fundamentales. La noción de "derechos de cuarta generación" se utiliza cada vez más para designar un haz de garantías vinculadas a la esfera digital, los datos y los algoritmos, que no sustituyen a los derechos clásicos, sino que los reformulan en un entorno de decisiones automatizadas. En el ámbito sanitario, estos derechos tienen una doble dimensión: la del paciente y la del profesional. Si el siglo XX consagró el derecho a la protección de la salud como derecho social, el siglo XXI exige reforzar la "dignidad algorítmica" de quienes reciben cuidados y de quienes los prestan. (Trujillo, 2025).

Desde la teoría de los derechos humanos se ha propuesto agrupar bajo esta cuarta generación a los llamados derechos "digitales": sobre los datos, la identidad digital, la explicación de las decisiones automatizadas y sobre la protección frente a formas nuevas de vigilancia y discriminación. En medicina, este bloque se traduce en garantías concretas frente a sistemas de IA capaces de influir de forma decisiva en el diagnóstico, el pronóstico o la priorización de recursos. La cuestión ya no es solo qué derechos asisten al paciente o al médico, sino cómo un modelo algorítmico puede erosionarlos de manera silenciosa si no se fijan salvaguardas jurídicas y organizativas adecuadas.

Los derechos de cuarta generación incluyen un **derecho a la alfabetización algorítmica y a la formación continua en IA**. La utilización responsable de sistemas complejos exige que el profesional comprenda sus sesgos, su comportamiento fuera de distribución y la forma adecuada de documentar discrepancias entre la recomendación automatizada y el juicio clínico. La propia AI Act reconoce la necesidad de garantizar niveles suficientes de "AI literacy" entre quienes operan sistemas de alto riesgo, lo que implica formación previa, acreditada y con tiempo protegido para los profesionales sanitarios. La ausencia de esa formación no puede trasladarse como culpa al médico individual, sino que debe considerarse un déficit estructural atribuible a la organización que despliega la tecnología.

Hay que señalar que se configura también un **derecho a un reparto equitativo de la responsabilidad jurídica** en contextos mediados por IA. Diversas propuestas doctrinales han mostrado que la introducción de

sistemas autónomos u opacos plantea tensiones importantes en el derecho de daños y en la responsabilidad profesional médica. Frente a la tentación de convertir al médico en "chivo expiatorio" de fallos sistémicos, los derechos de cuarta generación exigen mecanismos que vinculen la responsabilidad al agente causal: fabricante (cuando el daño deriva de un defecto de producto o de software), desplegador (cuando se trata de una implementación negligente, sin auditoría ni supervisión) u organización sanitaria (cuando no proporciona recursos, formación o circuitos de notificación de incidentes). El médico conserva su deber de diligencia, pero no debe cargar en solitario con las consecuencias de fallos de diseño o de gobernanza.

Puede proponerse un **derecho a la protección de los datos profesionales y a la no instrumentalización punitiva de la trazabilidad algorítmica**. Los registros de uso creados para garantizar seguridad del paciente y calidad asistencial no deberían convertirse en herramientas opacas de vigilancia laboral o de evaluación exclusivamente cuantitativa de la actuación clínica. La misma lógica que protege a los pacientes frente a decisiones totalmente automatizadas (por ejemplo, bajo el paraguas del RGPD) debe extenderse para evitar decisiones laborales o reputacionales de gran impacto basadas únicamente en perfiles generados por sistemas de IA sin garantías de revisión humana significativa.

Por último, los derechos de cuarta generación en medicina incluyen un **derecho a la participación en la gobernanza de la IA y a la libertad académica y de alerta en materia de seguridad**. Los profesionales sanitarios deben poder informar de sesgos, fallos o incidentes relacionados con sistemas de IA, así como publicar resultados de evaluación independiente, sin temor a represalias contractuales o disciplinaria. Esta dimensión enlaza con la tradición de la seguridad del paciente y la cultura de notificación de eventos adversos, y la proyecta en un ecosistema donde los algoritmos son ya actores de pleno derecho en los procesos clínicos.

En conjunto, la introducción de los derechos de cuarta generación permite formular un marco normativo en el que la protección de los pacientes y la protección de los médicos dejan de ser polos opuestos. La IA puede contribuir a reducir inequidades, mejorar diagnósticos y liberar tiempo clínico, pero solo si se articula sobre tres pilares: transparencia útil para decidir, gobernanza clínica robusta —con autonomía profesional reforzada— y responsabilidad proporcionada entre todos los actores del ecosistema algorítmico. El objetivo último es un **derecho a la salud aumentada** que

preserve, y no diluya, el pulso humano de la relación médico-paciente. (Razmetaeva, 2022).

2. Irrupción de nueva normativa europea y española con la ley de salud digital

La entrada en vigor del Reglamento (UE) 2024/1689, conocido como **Reglamento de Inteligencia Artificial (AI Act)**, el 1 de agosto de 2024, ha convertido a la Unión Europea en el principal polo regulatorio mundial en materia de IA, más aún tras la revocación en 2025 de la Orden Ejecutiva 14110 en Estados Unidos, que había constituido el marco de referencia norteamericano en este ámbito. Para la profesión médica, este desplazamiento del centro normativo significa que el ejercicio clínico en Europa va a estar crecientemente condicionado por un "bloque digital" de normas de obligado cumplimiento: al Reglamento de productos sanitarios (MDR) y de diagnóstico in vitro (IVDR), se suman ahora el AI Act, el Reglamento del Espacio Europeo de Datos de Salud (EEDS) y, en el caso español, una futura **Ley de Salud Digital** que adaptará el sistema sanitario nacional a este nuevo ecosistema jurídico. (Van Leeuwen et al, 2025).

En este contexto, los sistemas de IA utilizados en medicina —particularmente los que se integran en productos sanitarios de software— pasan a considerarse, con carácter general, **sistemas de alto riesgo**, lo que sitúa a los médicos y a las organizaciones sanitarias en la categoría de desplegadores (deployers) con obligaciones específicas. Estas no solo afectan a gestores y servicios de tecnologías de la información, sino también al clínico a pie de consulta, que es quien toma decisiones en última instancia. De manera indirecta, el AI Act transforma deberes de diligencia profesional —como la prudencia diagnóstica o la verificación crítica de resultados— en **exigencias jurídicas explícitas**, con impacto directo sobre la autonomía, la responsabilidad y las condiciones de trabajo del médico.

Uno de los elementos más novedosos es la consagración de la **alfabetización en IA (AI literacy)** como requisito legal. El artículo 4 del AI Act obliga a proveedores y desplegadores a garantizar un "nivel suficiente de alfabetización en IA" entre el personal que utiliza estos sistemas. Traducido al terreno profesional, esto implica que los médicos tienen derecho a

recibir formación específica y continuada en IA para poder comprender las capacidades y límites de los sistemas, interpretar sus salidas, detectar fallos y sesgos, y mantener un control humano significativo sobre la decisión clínica. La literatura científica ha señalado que, sin esta capacitación, el riesgo de **sesgo de automatización** y de delegación acrítica en la máquina aumenta de forma sustancial. Así, un deber normativo para la organización (formar a su personal) se convierte, desde la perspectiva de los derechos profesionales, en un **derecho del médico a ser formado adecuadamente** antes de verse obligado a trabajar con sistemas de IA de alto riesgo.

En paralelo, el AI Act refuerza la **supervisión humana y la trazabilidad**. El artículo 26 obliga a los desplegadores de sistemas de alto riesgo a establecer flujos de supervisión, asegurar que el personal pueda interrumpir o no seguir las recomendaciones de la IA y conservar los registros generados por el sistema durante, al menos, seis meses. Esto introduce, de facto, un derecho del médico a **no quedar subordinado a la decisión algorítmica** y a disponer de información suficiente para justificar por qué sigue o se aparta de una recomendación automatizada. La trazabilidad de logs y auditorías —reclamada por distintos autores en el ámbito de la medicina digital— refuerza la posibilidad de reconstruir el razonamiento asistido por IA y de distribuir de forma más justa las responsabilidades entre desarrolladores, instituciones y profesionales.

La transformación normativa europea no se limita a la lógica de riesgo de la IA. Con el **Reglamento (UE) 2025/327 sobre el Espacio Europeo de Datos de Salud (EEDS)**, en vigor desde marzo de 2025, la UE crea un marco común para el acceso, intercambio e interoperabilidad de los datos de salud electrónicos, tanto para uso asistencial como para investigación y políticas públicas. El objetivo declarado es mejorar el control de los ciudadanos sobre sus datos y, al mismo tiempo, facilitar el uso secundario con fuertes garantías de seguridad y gobernanza. Para los médicos, este Reglamento se traduce en la expectativa —y, por tanto, en un derecho exigible frente a las administraciones— de trabajar con **historias clínicas electrónicas interoperables**, completas y accesibles en el punto de atención, evitando la fragmentación de la información que hoy compromete la continuidad asistencial y la seguridad del paciente.

España ha comenzado a adaptar su ordenamiento a este nuevo escenario mediante el **Anteproyecto de Ley de Salud Digital**, actualmente en fase de consulta pública. Esta norma pretende regular la historia clínica digital interoperable a nivel nacional, el uso de tecnologías digitales —incluida

la IA— en la asistencia sanitaria, y la gobernanza del Espacio Europeo de Datos de Salud en nuestro país. Según los documentos preliminares del Ministerio de Sanidad, la futura ley definirá las funciones de las comunidades autónomas como autoridades regionales de salud digital, establecerá un modelo de gobernanza del uso primario y secundario de los datos y fijará las obligaciones de las administraciones para garantizar el acceso efectivo de pacientes y profesionales a la información clínica en cualquier punto del Sistema Nacional de Salud.

Desde la óptica de los derechos del médico, esta futura Ley de Salud Digital debería leerse como la concreción nacional de varios derechos profesionales emergentes en la era de la IA: el derecho a **trabajar con infraestructuras digitales seguras e interoperables**, el derecho a **claridad normativa** sobre el uso secundario de los datos que el propio médico contribuye a generar, y el derecho a que la digitalización no incremente de manera desproporcionada la carga burocrática ni deteriore la relación clínica. El alineamiento entre AI Act, EEDS, GDPR y Ley de Salud Digital constituye, en la práctica, una suerte de "constitución digital" de la práctica médica, que redefine las condiciones de ejercicio profesional y obliga a los colegios, sociedades científicas y órganos reguladores a reposicionar el catálogo de **derechos y deberes del médico** en un entorno crecientemente algorítmico.

En síntesis, la irrupción de esta nueva normativa europea y española no es solo un fenómeno técnico-jurídico; supone un cambio de paradigma para la profesión médica. El médico deja de ser un mero usuario pasivo de tecnologías digitales para convertirse en un sujeto jurídico central en su gobernanza: responsable de su uso diligente, pero también titular de derechos específicos a formación, información, interoperabilidad y protección frente a la automatización opaca. Los apartados siguientes explorarán cómo este nuevo marco se traduce en **derechos concretos del médico** en la práctica clínica cotidiana.

3. La figura clave del desplegador (deployer) en el AI Act.

La gran novedad del AI Act respecto a otros marcos tecnológicos es que no se limita a regular a quienes desarrollan o comercializan sistemas de IA (providers), sino que introduce expresamente la figura del "deployer" —que en este capítulo traducimos como desplegador sanitario— como sujeto regulado. Esto significa que el uso clínico de la IA deja de ser un espacio neutro: pasa a estar jurídicamente configurado y sometido a obligaciones propias, con impacto directo en la práctica médica cotidiana.

El artículo 3.4 del Reglamento (UE) 2024/1689 define al desplegador como toda persona física o jurídica, autoridad pública, agencia u otro organismo que utiliza un sistema de IA bajo su autoridad, salvo cuando lo haga en el marco de una actividad puramente personal o no profesional.

Esta definición es deliberadamente amplia y captura prácticamente a cualquier organización sanitaria que incorpore IA en sus procesos asistenciales o de gestión: hospitales, servicios de salud autonómicos, clínicas privadas, aseguradoras sanitarias, centros de diagnóstico, pero también, en determinados supuestos, un servicio clínico o incluso un profesional que decide poner en marcha y gestionar un sistema de IA relevante bajo su ámbito de decisión.

En la terminología del propio Reglamento, el desplegador es además una de las figuras incluidas en el concepto más amplio de operador, junto con el proveedor, el importador, el distribuidor y el fabricante de productos que integran IA. Esta arquitectura de "cadena de valor" permite repartir obligaciones, pero también prevé que, en determinadas circunstancias (por ejemplo, cuando se modifica sustancialmente el sistema o se le reasigna una finalidad de alto riesgo), el desplegador pase a ser considerado proveedor y asuma las cargas más intensas de ese rol.

Las obligaciones nucleares del desplegador se concentran en los sistemas de alto riesgo, entre los que se incluyen la mayoría de aplicaciones de IA sanitaria dirigidas al diagnóstico, triaje, planificación terapéutica o soporte a decisiones clínicas, especialmente cuando se integran en productos regulados como dispositivos médicos (MDR/IVDR).

El artículo 26 del AI Act fija un catálogo de deberes que puede agruparse en tres bloques:

1. **Uso conforme y medidas organizativas internas:**

    - Adoptar medidas técnicas y organizativas para garantizar que el sistema se utiliza de acuerdo con las instrucciones de uso facilitadas por el proveedor.
    - Asignar la supervisión humana a personas con la competencia, formación y autoridad necesarias, proporcionando además los apoyos materiales y de tiempo para que esta supervisión sea real y efectiva (no meramente nominal).
    - Implantar políticas de AI literacy (alfabetización en IA) para el personal que interactúe con el sistema, de forma proporcional al riesgo.

2. **Gestión de datos, funcionamiento y registro:**

    - Garantizar que los datos de entrada bajo su control sean pertinentes, suficientemente representativos y de calidad, de acuerdo con las especificaciones del sistema.
    - Monitorizar el funcionamiento del sistema de forma continua, detectando degradaciones de rendimiento, sesgos o comportamientos anómalos, e interrumpir su uso cuando exista un riesgo significativo para la seguridad o los derechos de las personas.
    - Conservar los registros (logs) generados automáticamente por el sistema, en la medida en que estén bajo su control, durante el periodo mínimo previsto por el Reglamento y la normativa sectorial aplicable.
    - Verificar, cuando proceda, que el sistema está correctamente inscrito en la base de datos europea de sistemas de IA de alto riesgo y abstenerse de utilizarlo si no cumple los requisitos de registro.

3. **Transparencia, impacto en derechos y cooperación regulatoria:**

    - Informar a las personas afectadas cuando se utilicen determinados sistemas de alto riesgo que tomen o apoyen decisiones con efectos significativos sobre ellas, obligación

que se añade a los deberes de información propios del RGPD.
- Realizar, en los supuestos previstos en el artículo 27, una evaluación de impacto sobre los derechos fundamentales previa al despliegue de determinados sistemas de alto riesgo, especialmente cuando los utilicen autoridades públicas o se afecte de forma sistemática a personas vulnerables.
- Coordinar estas evaluaciones con las evaluaciones de impacto en protección de datos (DPIA) exigidas por el RGPD, lo que en la práctica obliga a diseñar marcos integrados de gobernanza de datos y de IA en las organizaciones sanitarias.
- Notificar incidentes graves y no conformidades al proveedor y, en su caso, a las autoridades de vigilancia de mercado y de protección de datos, cooperando en la adopción de medidas correctoras.

Estas obligaciones son "sin perjuicio" de otras que el desplegador ya asume en virtud de la legislación sanitaria, de productos sanitarios, de protección de datos o de seguridad del paciente, lo que refuerza su posición como nodo central de responsabilidad en el ecosistema de IA clínica.

En el sector salud, la figura del desplegador tiene una traducción inmediata: los hospitales y servicios sanitarios que introducen IA en la práctica clínica se convierten en sujetos obligados, y los profesionales que trabajan en ellos —especialmente quienes lideran servicios, comités de innovación o direcciones médicas— pasan a desempeñar un papel activo en la gobernanza de esos sistemas.

La literatura reciente sobre la aplicación del AI Act en salud subraya tres consecuencias principales para médicos y organizaciones sanitarias:

- **Reconfiguración del deber de cuidado:** el estándar de buena práctica ya no se limita a "usar la mejor evidencia disponible", sino también a utilizar de forma diligente los sistemas de IA: entender sus indicaciones y contraindicaciones, supervisar su rendimiento y saber cuándo ignorar o desactivar una recomendación automatizada en favor del juicio clínico.

- **Necesidad de políticas internas de IA:** los centros deben dotarse de comités, protocolos y circuitos de evaluación que permitan cumplir con las obligaciones del desplegador (evaluaciones de impacto, registro y trazabilidad, formación, revisión periódica de algoritmos), integrando el AI Act con el MDR/IVDR y la normativa de historia clínica electrónica.

- **Refuerzo de los derechos del profesional:** para que el médico pueda asumir razonablemente estas obligaciones, necesita a su vez ciertos derechos: acceso a la documentación técnica relevante, formación específica en la herramienta, posibilidad efectiva de supervisión y override, y protección frente a responsabilidades desproporcionadas cuando el sistema no cumple los requisitos impuestos al proveedor.

La incorporación del desplegador como figura regulada desplaza el centro de gravedad de la responsabilidad desde un modelo centrado exclusivamente en el fabricante hacia un modelo compartido a lo largo de la cadena de valor. En medicina, esto significa que la organización sanitaria y, en último término, los profesionales que utilizan la IA en el cuidado de los pacientes no son meros "usuarios pasivos", sino operadores jurídicamente configurados, con obligaciones específicas pero también con la legitimidad para exigir garantías, transparencia y condiciones de uso que respeten la seguridad del paciente y los derechos fundamentales (García Luengo, 2025).

Por último, la configuración del médico y de la organización sanitaria como deployers de sistemas de alto riesgo tiene consecuencias directas sobre el seguro de responsabilidad civil profesional (RCP). El AI Act no regula de forma expresa el aseguramiento, pero reordena el mapa de riesgos y, con ello, la manera en que las pólizas deberán estructurarse para cubrir adecuadamente la práctica clínica asistida por IA.

El uso de sistemas de IA de alto riesgo en este tiempo pasa a formar parte del riesgo típico asegurado: diagnosticar, indicar pruebas o proponer tratamientos apoyándose en algoritmos deja de ser un elemento excepcional y se integra en la "lex artis digital". Sin embargo, al atribuir al deployer obligaciones específicas (supervisión humana, uso conforme a instrucciones, monitorización, conservación de logs, evaluación de impacto, etc.), se abre un espacio para que las aseguradoras diferencien entre:

- Riesgo asegurado por mala praxis clínica (errores de indicación, interpretación o seguimiento en el uso de la herramienta), y
- Riesgo no asegurado o limitado por incumplimiento organizativo (falta de formación, ausencia de supervisión humana real, uso fuera de las instrucciones del proveedor, omisión de evaluaciones de impacto o de notificación de incidentes.

En la práctica, esto puede traducirse en varios ajustes contractuales:

1. Cláusulas específicas sobre uso de IA. Muchas pólizas de RCP empezarán a incluir cláusulas que:

    - Exijan que los sistemas de IA utilizados estén debidamente certificados/registrados y se empleen en el marco de la indicación autorizada.
    - Subordinen la cobertura plena al cumplimiento razonable de las obligaciones del deployer (protocolos internos, supervisión humana, documentación de decisiones, conservación de logs).
    - Delimiten qué se considera "error médico" y qué se considera "fallo del sistema" atribuible al proveedor, abriendo la puerta a acciones de repetición o coaseguro entre aseguradoras de profesionales, centros y fabricantes.

2. Reforzamiento del componente organizativo. La responsabilidad se desplaza parcialmente del profesional aislado a la organización desplegadora (servicio, hospital, red sanitaria). Esto hace más relevante:

    - La existencia de pólizas colectivas de RCP que cubran tanto a la institución como a los profesionales, con límites y franquicias adaptados al nuevo riesgo tecnológico.
    - La evaluación de la gobernanza de la IA como criterio técnico de suscripción: comités de IA, protocolos de validación clínica local, auditorías periódicas de rendimiento y sesgos, planes de contingencia y cese de uso ante incidencias. Las organizaciones con mejor gobernanza podrán negociar condiciones más favorables (primas, límites, exclusiones).

3. Documentación, trazabilidad y prueba pericial. La obligación del deployer de conservar logs y registrar las interacciones con el sistema tiene un efecto inmediato en el terreno probatorio:

    - En un eventual procedimiento judicial, estos registros serán piezas clave para reconstruir si el médico actuó conforme a la lex artis y si supervisó adecuadamente la recomendación algorítmica.
    - La ausencia de registros o la imposibilidad de reconstruir el flujo de decisión puede interpretarse como indicio de incumplimiento de las obligaciones del deployer, lo que tensiona la posición del asegurado frente a su compañía.
    - A la vez, esta trazabilidad permite delimitar mejor la participación causal del sistema de IA, facilitando la pericia forense y el reparto de responsabilidad entre médico, centro y proveedor del sistema.

4. Nuevos escenarios de corresponsabilidad y cobertura cruzada El AI Act refuerza un modelo de responsabilidad en cadena, donde proveedor, deployer y, eventualmente, otros operadores pueden compartir responsabilidad por un mismo daño. Ello apunta a escenarios de:

    - Pólizas coordinadas (o incluso productos específicos de “RCP + ciber/IA”) que combinen responsabilidad profesional, riesgo tecnológico y ciberseguridad.
    - Mayor litigiosidad entre aseguradoras para determinar el porcentaje de responsabilidad imputable a defecto del sistema (producto) frente a error de uso (servicio médico), lo que incentivará la precisión en las cláusulas de cobertura y exclusión.

Desde la perspectiva del médico individual, la condición de deployer no debe entenderse solo como una carga añadida, sino también como un argumento para reivindicar derechos aseguradores: acceso a información técnica suficiente, formación acreditada, protocolos claros de uso y posibilidad de documentar adecuadamente la supervisión clínica. Sin estas condiciones, el profesional queda expuesto a un riesgo jurídico y asegurador desproporcionado.

En definitiva, la irrupción del deployer en el AI Act obliga a repensar la RCP como un instrumento no solo de reparación, sino de gobernanza preventiva de la IA clínica: las pólizas pasan a ser un incentivo más para que hospitales y médicos integren la gestión del riesgo algorítmico en su práctica diaria, alineando cumplimiento normativo, seguridad del paciente y protección efectiva del profesional (Van Leeuwen et al, 2025).

4. La nueva propuesta europea de simplificación de la Ley Ómnibus digital (AI Act, GDPR, e-Privacy, Data Act).

El 19 de noviembre de 2025 la Comisión Europea presentó un paquete de simplificación de legislación digital —conocido como Digital Omnibus o Digital Package— que afecta de forma coordinada al AI Act y a otras normas clave (GDPR, e-Privacy, Data Act). La Comisión lo justifica como una medida para reducir cargas administrativas, mejorar competitividad y evitar una "fuga de innovación", en un escenario de presión industrial y geopolítica evidente.

En lo relativo al AI Act, no estamos ante un simple ajuste técnico, sino ante una reprogramación del despliegue efectivo de su núcleo regulatorio: las obligaciones exigibles a los sistemas de alto riesgo aplicables en sectores sensibles como la salud (White & Case, 2025).

7 Derechos y obligaciones del médico ante la IA

Dr. José Antonio Trujillo

La propuesta introduce tres movimientos con gran relevancia sanitaria:

1. **Retraso del régimen de alto riesgo (Anexo III) hasta finales de 2027.**
   Las obligaciones completas para sistemas de alto riesgo —entre ellos los utilizados en salud (diagnóstico, triaje, decisiones terapéuticas asistidas, etc.)— pasan de la fecha prevista (2 de agosto de 2026) a un "longstop" en **diciembre de 2027**, condicionado además a que la Comisión declare listos los estándares y guías técnicas.

2. **Aplicación todavía más tardía para algunos sistemas encuadrados en Anexo I.**
   En sectores ligados a producto regulado (p. ej., software como producto sanitario), el Omnibus abre la puerta a que ciertas obligaciones arranquen hasta **agosto de 2028** si se demora la decisión de activación.

3. **Simplificación y proporcionalidad administrativa.**
   Se plantean ajustes de trámite (menos obligaciones de registro público cuando el proveedor documenta que el uso no es alto riesgo, y un régimen sancionador más proporcional para empresas medianas). Aunque se presenta como "clarificación", implica menor presión formal en la fase inicial de cumplimiento.

Otro tema clave es cómo queda la figura del deployer (usuario profesional que despliega IA en práctica clínica o institucional). No desaparece del AI Act, pero la propuesta **sí retrasa su aterrizaje práctico**. El motivo es simple: los deberes más característicos del deployer nacen precisamente del régimen de alto riesgo, que queda aplazado:

- **Antes del Omnibus:** el deployer en salud iba a verse obligado, desde agosto de 2026, a asumir un bloque completo de responsabilidades exigibles y sancionables.

- **Con el Omnibus:** ese bloque se posterga 16–18 meses hasta diciembre de 2027, reduciendo el incentivo inmediato a invertir en estructuras internas de cumplimiento.

- **En la práctica, entre 2025 y 2027 conviviremos con un deployer "ligero":** reconocido normativamente, pero sin el peso completo

del compliance clínico-tecnológico que el legislador había diseñado.

Esta nueva propuesta comunitaria afecta sobre todo a las obligaciones que exigen cultura organizativa, inversión técnica y trazabilidad clínica:

- **Uso conforme a finalidad prevista y control de calidad de datos de entrada.**
  El deployer debe garantizar que el sistema se usa en el contexto clínico autorizado, con datos adecuados y sin desbordar indicaciones. Este control es costoso y su urgencia se diluye con la prórroga.

- **Supervisión humana reglada.**
  La supervisión no es solo "tener un médico delante", sino diseñar protocolos para intervenir, anular o revertir salidas del sistema. Esa formalización queda pospuesta.

- **Monitorización post-despliegue, logging e informe de incidentes.**
  En sanidad es la pieza más crítica: detectar drift, errores clínicos, sesgos emergentes y reportar eventos adversos. El Omnibus desplaza la obligación plena de monitorizar y notificar.

- **Auditoría interna y trazabilidad demostrable.**
  El sistema probatorio de cumplimiento (documentación, evaluaciones internas periódicas) se convierte, durante dos años más, en un terreno cuasi-voluntario.

No todo queda en suspenso. Hay deberes ya vigentes o de activación temprana que mantienen viva la figura:

- **Respeto a prohibiciones y límites estructurales del AI Act.**
  El deployer no puede usar sistemas prohibidos ni prácticas que vulneren derechos fundamentales. Esta obligación rige ya.

- **Alfabetización en IA (AI literacy).**
  Aunque el Omnibus tiende a suavizar su exigibilidad estricta, sigue marcando una expectativa clara de formación continuada y cultura de uso responsable, especialmente en sectores sensibles.

En el entorno asistencial, esto se traduce en dos tareas inmediatas: seleccionar tecnologías con prudencia y formar a los clínicos para entender límites, sesgos y condiciones de uso (Bignami et al, 2025).

La simplificación legal genera un efecto dual:

Aspecto protector:

El deployer gana margen frente a un cumplimiento prematuro sin estándares técnicos cerrados. Es un argumento fuerte de seguridad jurídica: no se puede exigir lo que aún no está técnicamente definido.

Aspecto problemático:

Aplaza la consolidación de sus derechos operativos: derecho a guías claras, a canales estables de reporte, a criterios uniformes de auditoría y a un ecosistema de certificación maduro. El deployer queda más tiempo navegando entre incertidumbre normativa y presión asistencial.

En sanidad, la prórroga no es neutra. Abre tres riesgos que justifican una reacción profesional proactiva:

1. **Ventana regulatoria amplia hasta 12/2027.** Habrá más tiempo para desplegar IA clínica sin obligaciones plenas de monitorización y reporte. Eso aumenta la probabilidad de "deuda de cumplimiento" acumulada cuando llegue la fecha dura.
2. **Asimetría entre grandes proveedores y deployers locales.** Los proveedores globales seguirán adaptando productos a compliance futuro; muchas clínicas y servicios hospitalarios tenderán a esperar. Resultado: adopción rápida con gobernanza lenta.
3. **Confusión cultural entre simplificación y rebaja de garantías.** Parte de la crítica europea alerta de un retroceso real en protección digital. Si esa percepción se instala, se erosiona la confianza pública en IA clínica y se polariza el debate médico-social.

El retraso, paradójicamente, abre una ventana útil: **dos años para que la medicina se anticipe**.
Colegios profesionales, sociedades científicas y servicios de salud pueden ocupar el vacío con:

- Guías clínicas de despliegue seguro.
- Estándares de supervisión humana y trazabilidad adaptados a cada especialidad.
- Protocolos de monitorización e incidente clínico-algorítmico.
- Modelos de contrato y gobernanza.

Si no lo hacemos, cuando llegue diciembre de 2027 el deployer sanitario tendrá que asumir obligaciones plenas de golpe, con sistemas ya implantados y sin cultura previa. Si lo hacemos, llegaremos a esa fecha con una figura madura, deontológicamente alineada y clínicamente segura.

5. Derechos fundamentales de los médicos frente a la IA.

La incorporación de la IA de alto riesgo a la práctica clínica no solo genera nuevas obligaciones para el médico como desplegador (deployer) (Osborne, 2024), sino también un haz de derechos fundamentales sin los cuales su responsabilidad sería materialmente imposible de ejercer. A continuación, se desarrollan, en clave jurídica y práctica, los principales derechos del profesional sanitario frente a la IA (Funner et al, 2024) :

1.1. **Autonomía clínica reforzada**

La autonomía clínica es el derecho del médico a tomar la decisión final sobre el diagnóstico y el tratamiento, asistido pero nunca sustituido por la tecnología. El AI Act exige que los sistemas de alto riesgo permitan una supervisión humana efectiva, incluyendo la posibilidad de intervenir, modificar o interrumpir su funcionamiento para evitar daños a la salud o a los derechos fundamentales.

El **Código Internacional de Ética Médica de la WMA** refuerza esta idea al exigir que el médico preserve siempre su juicio profesional

independiente, incluso frente a presiones institucionales o tecnológicas.

En términos prácticos, este derecho implica:

- Posibilidad técnica y organizativa de desoír, desactivar o revertir una recomendación algorítmica (el "botón de parada clínico").
- Protección deontológica y contractual frente a represalias cuando el médico se aparta justificadamente de la recomendación de la IA.
- Reconocimiento de que la IA es una segunda opinión ampliada, no un oráculo vinculante.

Ejemplos:

- Real: Un algoritmo de clasificación de lesiones cutáneas etiqueta un lunar como "benigno con un 95 % de confianza". El dermatólogo decide extirparlo porque ha observado cambios recientes y características clínicas sutiles no captadas por el modelo. La biopsia confirma un melanoma inicial: la decisión autónoma del profesional evita un falso negativo con consecuencias graves.
- Futuro: Un algoritmo de gestión de camas de UCI recomienda el alta de un paciente cardiaco porque los parámetros monitorizados se han normalizado. El cardiólogo detecta astenia marcada y un empeoramiento subjetivo que no aparece en los registros. Mantiene al paciente ingresado; horas después, el enfermo desarrolla un cuadro de insuficiencia cardiaca incipiente.

1.2. Debida diligencia profesional (lex artis) y "última palabra"

La **lex artis ad hoc** se actualiza para incorporar el uso razonable de la IA, pero no se sustituye por ella. El médico tiene el derecho-deber de usar la IA con la misma exigencia crítica que aplicaría a cualquier prueba diagnóstica, lo que incluye conocer sus limitaciones, la población en la que fue validada y las condiciones de uso seguro.

La experiencia nos dice que debemos ser siempre prudentes en la adopción de las aplicaciones médicas que utilizan IA, porque sus expectativas pueden ser finalmente poco realistas. Así por ejemplo, la literatura sobre sepsis y sistemas de alerta electrónicos ha mostrado cómo modelos ampliamente

difundidos, como el **Epic Sepsis Model**, presentaban baja discriminación y mala calibración en validaciones externas, obligando a mantener un juicio clínico independiente y a no "delegar" ciegamente en la herramienta. De forma similar, la experiencia de **Watson for Oncology** en MD Anderson, con recomendaciones oncológicas inseguras, ilustra que un sistema prestigioso puede fallar gravemente y ser retirado.

Este derecho se concreta en:

- Mantener la **última palabra clínica**, aunque exista una recomendación algorítmica.
- Exigir evidencia científica y validación externa antes de incorporar un algoritmo a la práctica habitual.
- Rehusar el uso de una herramienta cuando no exista información suficiente sobre su rendimiento en el contexto local.

Ejemplos:

- Real: Un radiólogo se limita a aceptar el informe de un algoritmo de detección de nódulos pulmonares sin revisar íntegramente la tomografía. Un pequeño nódulo maligno, visible para un especialista diligente, pasa inadvertido. En un eventual procedimiento, un tribunal podría considerar que el radiólogo vulneró la lex artis al depender exclusivamente de la IA pese a disponer de los cortes completos.
- Futuro: En urgencias, una app de triaje basada en IA clasifica a un paciente con dolor abdominal atípico como "baja prioridad". El médico, cumpliendo su deber de diligencia, realiza anamnesis y exploración completas, detecta signos de irritación peritoneal y decide ingreso para descartar una apendicitis retrocecal. La actuación correcta se atribuye al profesional; la app solo era un insumo más.

1.3. **Condiciones de trabajo seguras (IA validada y mantenida)**

El derecho del médico a un entorno de trabajo seguro, reconocido por la **Directiva 89/391/CEE** y por la **Ley 31/1995 de Prevención de Riesgos Laborales**, incluye también la protección frente a tecnologías inseguras o no validadas.
En IA sanitaria, esto implica que el empleador y el desplegador (deployer) deben garantizar que:

- Los sistemas de IA que actúan como **Software as a Medical Device (SaMD)** cumplen el Reglamento (UE) 2017/745 (MDR): evaluación clínica, gestión de riesgos, vigilancia poscomercialización.
- Existen procesos formales de **validación local**, detección de drift (deriva) y mantenimiento correctivo antes y durante su uso.
- Se puede suspender el sistema si se detectan problemas de seguridad, sin penalizar al profesional que lo reclama. El AI Act refuerza este derecho al exigir supervisión humana y registros de uso para sistemas de alto riesgo.

Ejemplos:

- Realista: Un hospital implanta un algoritmo de sepsis entrenado en otro país. Tras unas semanas, los clínicos detectan una avalancha de falsos positivos en pacientes con comorbilidades habituales en su población. La sobrecarga asistencial y la "fatiga de alarmas" aumentan el estrés y el riesgo de burnout. El servicio de Medicina Interna invoca su derecho a condiciones seguras para exigir la suspensión temporal del sistema y su revalidación con datos locales.
- Futuro: Un módulo de planificación de prótesis de rodilla basado en IA empieza a generar mediciones claramente inconsistentes tras una actualización de software. El traumatólogo se niega a utilizarlo hasta que el servicio técnico y el fabricante certifiquen su corrección. La institución debe respetar esa decisión y abstenerse de presionarle para "seguir como siempre".

1.4. Derecho a la formación continua y a la alfabetización en IA

El AI Act introduce una obligación específica de **alfabetización en IA**: proveedores y desplegadores (deployers) deben asegurar que el personal que utiliza estos sistemas tenga un nivel adecuado de conocimientos para usarlos de forma segura y responsable.

En España, la **Ley 44/2003 de ordenación de las profesiones sanitarias (LOPS)** define la formación continuada como un proceso permanente al que los profesionales tienen derecho y obligación, destinado a actualizar sus competencias ante la evolución científica y tecnológica. La confluencia de ambas normas respalda un derecho específico del médico a:

- Recibir **formación previa, estructurada, acreditada y remunerada** antes de verse obligado a usar IA de alto riesgo.

- Comprender los **principios básicos de funcionamiento** del algoritmo, sus sesgos conocidos, sus limitaciones y el significado de sus salidas (p.ej. scores, probabilidades, umbrales).
- Disponer de protocolos claros de override y reporte de incidentes asociados a la IA.

Ejemplos:

- Realista: Un sistema de IA para cribado de retinopatía diabética se despliega en Atención Primaria. Los médicos reciben solo un folleto de dos páginas sobre el uso de la interfaz. Nadie les explica que la sensibilidad del modelo es menor en determinados grupos étnicos subrepresentados en los datos de entrenamiento. El déficit formativo genera una falsa sensación de seguridad y varios falsos negativos.
- Futuro: Un hospital pone en marcha un sistema de apoyo a la prescripción de quimioterapia ajustada a biomarcadores. Como parte del proyecto, todos los oncólogos realizan un curso de 40 horas acreditado por la comisión de formación continuada, con tiempo protegido y simulaciones de casos. La implementación se acompaña de sesiones de morbi-mortalidad específicas para revisar los efectos del modelo y detectar desviaciones.

1.5. Responsabilidad proporcional y reparto justo del riesgo

El médico tiene derecho a no ser convertido en "chivo expiatorio" de fallos que son, en realidad, de diseño algorítmico, ciberseguridad o implementación organizativa. El nuevo **Producto Liability Directive (Directiva (UE) 2024/2853)** amplía la responsabilidad objetiva por productos defectuosos a software y sistemas de IA, incluyendo actualizaciones y fallos de ciberseguridad.

En paralelo, el AI Act coloca obligaciones específicas sobre proveedores y desplegadores (deployers) (registros, vigilancia, evaluación del riesgo), de modo que el daño derivado de un defecto del sistema no pueda recaer íntegramente sobre el profesional que lo utilizó conforme a la lex artis y a la formación recibida.

Este derecho se traduce en que:

- Los errores **atribuidos a defectos del producto** (p.ej. un bug que infradiagnostica sistemáticamente, datos de entrenamiento gravemente

sesgados, fallo de ciberseguridad) se ventilan prioritariamente en el marco de la responsabilidad por producto.
- La organización sanitaria responde por **incumplir sus deberes como deployer** (formación insuficiente, ausencia de monitorización, no realización de evaluaciones de impacto, falta de protocolos).
- El médico solo asume responsabilidad cuando incumple su deber de diligencia (por ejemplo, no revisa una imagen o no explora a un paciente porque "la IA ya lo ha hecho").

Ejemplos:

- Realista: Una actualización introduce un error en el software de IA que analiza TAC torácicos, reduciendo su sensibilidad para nódulos pequeños. El hospital no fue informado y siguió utilizando el sistema de buena fe. Tras varios retrasos diagnósticos, las reclamaciones deben dirigirse principalmente contra el fabricante por producto defectuoso y, en su caso, contra la administración que no implantó un régimen de vigilancia adecuado, no contra cada radiólogo individual.
- Futuro: Un sistema de bombas de insulina conectadas a IA recibe datos erróneos de un sensor defectuoso, provocando una hipoglucemia grave. El endocrinólogo había ajustado los parámetros conforme al protocolo y revisaba periódicamente las tendencias. La responsabilidad se focaliza en el fabricante del sensor y del software, apoyándose en la Directiva de productos defectuosos, no en el médico que confió razonablemente en el dispositivo.

1.6. Libertad académica e intelectual (alertas de seguridad y crítica pública)

Para que la IA mejore, los profesionales deben poder **investigar, documentar y comunicar** sus fallos, sesgos y riesgos sin temor a represalias. La **Directiva (UE) 2019/1937 sobre protección de informantes** y su trasposición en España mediante la **Ley 2/2023** protegen a quienes, en un contexto laboral, informan de infracciones normativas o riesgos graves para el interés público.

En paralelo, el MDR y el AI Act obligan a los operadores a reportar incidentes graves y problemas de seguridad, lo que incluye fallos relevantes de IA utilizada como producto sanitario.

Este derecho implica:

- Posibilidad de notificar internamente y a la autoridad competente los problemas detectados en un sistema de IA sin sufrir represalias disciplinarias o contractuales.
- Capacidad de publicar resultados científicos sobre el rendimiento real, los sesgos o las limitaciones de un sistema, siempre respetando la confidencialidad de pacientes y los canales de seguridad.
- Protección frente a cláusulas de confidencialidad abusivas que intenten impedir la comunicación de riesgos para la salud pública.

Ejemplos:

- Realista: Un dermatólogo detecta que el modelo de melanoma usado en su hospital infradiagnostica sistemáticamente lesiones en fototipos oscuros. Documenta una serie de casos, informa al comité de seguridad y plantea una publicación en una revista científica. La protección del informante y las obligaciones de vigilancia del MDR amparan su actuación, aunque el fabricante prefiera no divulgar esos datos.
- Futuro: Un equipo de cirujanos registra un patrón de errores de un robot quirúrgico en determinadas posiciones anatómicas. Redactan un informe técnico y un artículo científico para alertar a la comunidad. El hospital establece un canal seguro, al amparo de la Ley 2/2023, para que estos hallazgos se comuniquen sin riesgo de represalias ni por parte de la dirección ni del fabricante.

1.7. Protección de los datos personales y de las métricas del profesional

Los sistemas de IA generan de forma rutinaria **logs, métricas de rendimiento y patrones de uso** que pueden referirse tanto a pacientes como a profesionales. El AI Act exige conservar registros para garantizar trazabilidad y seguridad, pero también obliga a informar a los trabajadores cuando la IA se utiliza en el contexto laboral, especialmente si se trata de sistemas de alto riesgo para gestión de personas.

El **RGPD**, en su artículo 22, reconoce a las personas el derecho a no ser objeto de decisiones basadas únicamente en tratamiento automatizado con efectos jurídicos o similares, y permite establecer salvaguardas reforzadas en el ámbito laboral. La **LOPDGDD** española completa este marco con derechos específicos sobre el uso de dispositivos digitales, videovigilancia y

geolocalización en el trabajo, subrayando la necesidad de informar y limitar el tratamiento.

En consecuencia, el médico tiene derecho a que:

- Los datos de uso de la IA y sus métricas individuales no se conviertan en un sistema de **vigilancia encubierta** ni en base para sanciones automatizadas.
- Cualquier cuadro de mando que utilice indicadores derivados de IA para evaluar su desempeño esté sometido a **evaluación de impacto en protección de datos**, información previa clara y posibilidad real de impugnar decisiones.
- Se apliquen principios de minimización, anonimización o agregación cuando la finalidad sea la mejora de calidad o la investigación, y no el control individual del trabajador.

Ejemplos:

- Realista: Un hospital usa la información sobre cuántas ecocardiografías informa cada cardiólogo y cuántas veces acepta o rechaza la sugerencia de la IA para construir un ranking de productividad interno. El derecho a la protección de datos laborales y las limitaciones del artículo 22 RGPD exigen una evaluación de impacto, información transparente y la prohibición de decisiones automáticas sin intervención humana cualificada.
- Futuro: Un sistema registra cada discrepancia entre el informe del radiólogo y la recomendación algorítmica. La dirección plantea usar un "índice de discrepancia con la IA" como criterio de renovación contractual. Esta práctica podría vulnerar tanto el derecho a la autonomía clínica como las garantías frente a decisiones automatizadas, y justificaría la intervención de los delegados de protección de datos y de la representación profesional.

6. OBJECCIÓN DE CONCIENCIA E IA.

En el marco de la medicina digital, puede hablarse de **objeción de conciencia tecnológica** como el derecho del profesional sanitario a negarse, de forma motivada y documentada, a utilizar un sistema concreto de inteligencia artificial o de automatización clínica cuando considere que su uso compromete la **seguridad, la dignidad o la adecuada atención del paciente**.

No se trata de un rechazo genérico a la tecnología ni de una resistencia irracional al cambio, sino de una forma específica de objeción profesional que se sitúa en la intersección de (Herreros et al, 2024):

- La autonomía profesional e independencia clínica del médico.
- El deber de no maleficencia y de protección de la integridad del paciente, tal como reconoce el Convenio de Oviedo (primacía del ser humano frente al interés de la ciencia o la sociedad).
- El marco regulatorio de la IA, que exige supervisión humana eficaz sobre los sistemas de alto riesgo (AI Act, art. 14 y obligaciones del desplegador o deployer).

Desde la tradición bioética, la objeción de conciencia se ha defendido como un mecanismo para preservar la **integridad moral del profesional** en contextos como el aborto o la eutanasia. La novedad en el contexto de la IA no es el "derecho a objetar" en sí, sino **su objeto**: ya no se dirige únicamente contra determinados actos clínicos, sino contra el uso de **determinadas herramientas algorítmicas** cuando éstas se perciben como inseguras, sesgadas o incompatibles con la lex artis.

En el contexto español, esta figura puede interpretarse como una extensión de la llamada **"objeción de ciencia"**, ya introducida en el debate deontológico, entendida como el rechazo de una norma u orden por razones científicas o profesionales que se oponen a la libertad de método o de prescripción del médico. La objeción tecnológica sería, en este sentido, la aplicación de esa objeción de ciencia a sistemas de IA y tecnologías automatizadas.

6.2. Fundamento jurídico y ético

La objeción de conciencia tecnológica no nace en el vacío, sino que se apoya en varios pilares normativos y éticos ya consolidados (Celie et al, 2024):

1. **AI Act (Reglamento (UE) 2024/1689):**
    - El art. 14 exige que los sistemas de IA de alto riesgo permitan una **supervisión humana efectiva**, incluyendo la posibilidad de **no utilizar el sistema, ignorar, revertir o detener su salida** ("stop button") cuando el profesional lo considere oportuno para proteger la salud o los derechos fundamentales.
    - El art. 26 obliga al deployer a asignar a personas con la **competencia, formación y autoridad necesarias** la tarea de supervisar y, en su caso, corregir o suspender el uso de la IA.
    - La propia estructura del AI Act presupone que el médico no es un "operador pasivo", sino un **agente responsable** que puede y debe intervenir activamente ante riesgos.
2. **Convenio de Oviedo (1997)**
    - Art. 2: afirma la **primacía del ser humano** sobre el exclusivo interés de la sociedad o de la ciencia, estableciendo un límite a cualquier innovación biomédica que ponga en riesgo la dignidad o los derechos de la persona.
    - Art. 5 y ss.: refuerzan el rol del profesional como garante de que toda intervención se realice en condiciones de información, consentimiento y proporcionalidad.
3. **Deontología profesional (Código de Deontología Médica español)**
    - Reconoce el **derecho y deber de ejercer con autonomía profesional e independencia clínica**, así como la libertad de método y prescripción, y obliga a denunciar deficiencias que impidan una atención correcta.
    - Los desarrollos recientes del Código incluyen capítulos específicos sobre **objeción de conciencia** y aluden expresamente a la **objeción de ciencia**, lo que facilita conceptualmente la extensión a la tecnología.

4. **Ética profesional internacional (WMA, AMA y otros)**

    - La World Medical Association, en su declaración sobre inteligencia artificial en la atención médica, insiste en que los sistemas de IA deben **reforzar, y no sustituir**, el juicio clínico, y que los médicos deben poder rechazar tecnologías que no sean seguras, validadas o explicables.
    - La American Medical Association, en sus principios sobre "augmented intelligence", establece que los médicos deben mantener la **capacidad de revisar, cuestionar y apartarse** de las recomendaciones algorítmicas cuando entren en conflicto con el interés del paciente.

5. **Doctrina sobre objeción de conciencia en sanidad**

    - La literatura reciente subraya que, en sociedades plurales, la objeción de conciencia en salud es un derecho reconocido, aunque debe armonizarse con el acceso del paciente a prestaciones legales.
    - Parte de esta doctrina apunta precisamente a la expansión de la objeción hacia nuevas prácticas tecnológicamente complejas.

Sobre esta base, la objeción de conciencia tecnológica puede conceptualizarse como:

"El derecho del médico a abstenerse de utilizar un sistema de IA o automatización clínica cuando, en conciencia profesional y sobre bases científicas razonables, considere que su uso vulnera la lex artis, pone en riesgo injustificado al paciente o erosiona de manera sustantiva su autonomía clínica".


CRITICAL
ERROR

En ningún caso, la objeción de conciencia tecnológica no debe confundirse con el rechazo genérico a la digitalización. Su espacio natural es la **frontera entre el uso razonable y el uso imprudente** de la IA. Algunos escenarios paradigmáticos:

1. **Triaje automatizado con criterios injustificados:**
    - Un hospital implanta un sistema de triaje de urgencias que prioriza pacientes en función de variables sociodemográficas (código postal, aseguradora, nivel de renta indirecto) sin justificación clínica ni transparencia en el modelo.
    - Un médico de urgencias detecta que, de manera sistemática, las personas mayores que viven solas quedan en categorías de menor prioridad, pese a cuadros clínicos graves. Decide **no usar** el algoritmo en su turno y documenta en la historia que aplica triaje clínico tradicional, alegando riesgo de discriminación.
    - Aquí, la objeción se fundamenta en la **no maleficencia**, la igualdad de trato y la **prohibición de discriminación algorítmica** prevista en el AI Act.

2. **Sistema de ayuda al diagnóstico "caja negra" con falsos negativos:**
    - Un servicio de radiología adopta un sistema de detección de nódulos pulmonares con alta sensibilidad declarada. Estudios internos revelan, sin embargo, un número relevante de falsos negativos y el proveedor se niega a compartir las bases de validación o criterios de funcionamiento.
    - Un radiólogo puede alegar objeción de conciencia tecnológica y **exigir realizar su propia lectura sistemática**, utilizando la IA sólo de modo exploratorio o no utilizándola, hasta que se valide localmente el rendimiento.

3. **Modelos no adaptados al contexto local:**
    - Un algoritmo de predicción quirúrgica se ha entrenado en hospitales de alto volumen con pacientes de bajo riesgo y tecnología distinta. En un hospital comarcal, los resultados muestran discrepancias importantes (p.ej., infravalora complicaciones en población anciana con comorbilidades).

- El cirujano puede objetar el uso de ese modelo para la planificación en su centro hasta que se realice una **recalibración local** y se establezca una evidencia de validez externa, invocando la obligación de evitar prácticas clínicas no validadas.

4. **Automatización excesiva en salud mental:**

   - Una organización implanta un chatbot para el seguimiento de pacientes con ideación suicida, reduciendo significativamente las entrevistas presenciales.
   - El psiquiatra responsable puede oponerse a delegar en la IA la valoración de riesgo agudo, manteniendo la exigencia de **contacto humano directo** en pacientes de alto riesgo, y alegando que el modelo no garantiza la contención emocional ni la detección fina de señales de peligro.

En todos estos casos, la objeción no supone un rechazo absoluto de la IA como concepto, sino de su **uso concreto en condiciones que el médico considera profesionalmente inaceptables**.

Aunque el término "objeción de conciencia tecnológica" no está todavía positivizado como tal, pueden observarse tendencias convergentes en distintos entornos:

- Francia
  - Informes del Comité Consultivo Nacional de Ética (CCNE) sobre IA y salud subrayan que los sistemas algorítmicos deben permanecer "bajo control del médico" y que la autonomía profesional incluye la facultad de **rechazar herramientas que no sean suficientemente validadas o comprensibles**.
  - Se discute la necesidad de incorporar explícitamente en los códigos deontológicos la facultad del médico de **no utilizar sistemas que socaven su criterio clínico**.

- Canadá (Ontario)
  - Las guías del Colegio de Médicos y Cirujanos de Ontario sobre uso de IA en práctica clínica recalcan que el médico sigue siendo **plenamente responsable** de las decisiones, que debe entender las limitaciones de los sistemas y que

puede apartarse de sus recomendaciones cuando lo exija el interés del paciente, documentando su razonamiento.

- Estados Unidos (AMA)
    - La AMA, en sus políticas sobre "augmented intelligence in health care", reconoce que la IA debe **apoyar y no reemplazar** el juicio clínico, y que los médicos tienen la obligación –y en la práctica, el derecho– de **rechazar sistemas que no sean seguros, efectivos o transparentes**.

- Alemania y doctrina europea
    - La literatura doctrinal alemana y europea sobre IA sanitaria ha empezado a explorar figuras cercanas a la "technologische Gewissensverweigerung" (negativa de conciencia tecnológica), ligadas a la preservación de la responsabilidad profesional frente a entornos altamente automatizados.

Estas tendencias apuntan a un principio compartido: **el médico no puede ser obligado a utilizar una IA que considere incompatible con la buena práctica y la seguridad del paciente**, siempre que su negativa sea razonada, proporcionada y compatible con el acceso del paciente a la atención necesaria.

De cara a una regulación más explícita de la objeción de conciencia tecnológica en el entorno español, pueden proponerse varias vías complementarias:

1. **Nivel deontológico (Código de Deontología Médica):**

    - Introducir un artículo específico sobre **objeción de conciencia tecnológica o de ciencia aplicada a IA**, que:

        - Reconozca el derecho del médico a **no utilizar un sistema de IA** cuando existan motivos científicos o éticos fundados para considerar que su uso vulnera la lex artis o la seguridad del paciente.
        - Exija que la objeción esté **motivada, documentada y comunicada** al responsabl asistencial y, en su caso, al comité de ética del centro.

- Prohíba que esa objeción dé lugar a sanciones disciplinarias si se ejerce de buena fe y en defensa del paciente.

2. **Protocolos institucionales y comités de ética algorítmica:**

   - Incluir en los protocolos de despliegue de IA de alto riesgo una sección específica de **"override y objeción profesional"** que detalle:

     - Cómo desactivar o ignorar el sistema sin generar riesgos adicionales.
     - Cómo registrar en la historia clínica la decisión y sus motivos.
     - Cómo remitir casos recurrentes al **comité de ética asistencial o algorítmica** para evaluar si hay problemas sistémicos en la tecnología.

3. **Cláusulas contractuales y protección laboral:**

   - Incorporar en los contratos de médicos de centros públicos y privados una cláusula que:

     - Reconozca el derecho del profesional a **negarse a utilizar sistemas no validados, no explicables o no ajustados a la normativa** (AI Act, MDR, RGPD, etc.).
     - Garantice que el ejercicio razonable de esa negativa no puede ser causa de sanción, siempre que se asegure la continuidad asistencial por vías alternativas.

4. **Condiciones de ejercicio de la objeción:**

    - Para evitar abusos o disfunciones, la objeción de conciencia tecnológica debería:

        - Basarse en **razones profesionales serias**, no en comodidad u oportunismo.
        - Mantener la prioridad del **acceso del paciente a la atención**, facilitando, cuando sea posible, que otro profesional utilice la herramienta si es segura y está validada.
        - Ser **revisable**: un sistema inicialmente objetado puede dejar de serlo tras su adecuada validación, mejora o adaptación local.

En conclusión, podemos decir que la objeción de conciencia tecnológica constituye una pieza clave en la **gobernanza ética de la IA en medicina**. Lejos de bloquear la innovación, la encauza:

- Protege al médico como **último garante de la seguridad del paciente** frente a decisiones automatizadas opacas o insuficientemente validadas.
- Refuerza el mandato del AI Act de mantener una **supervisión humana efectiva**, al reconocer que esa supervisión incluye el derecho a decir "no" cuando el sistema entra en conflicto con la buena práctica.
- Conecta la tradición de la **objeción de conciencia y de ciencia** en medicina con los desafíos de la era algorítmica, evitando que la IA erosione silenciosamente la autonomía profesional.

En un ecosistema sanitario crecientemente automatizado, garantizar que el médico pueda **interrumpir, cuestionar o rechazar** una IA concreta –sin represalias, con responsabilidad y transparencia– es una condición imprescindible para que la innovación tecnológica sea compatible con la ética clínica, la dignidad del paciente y los derechos fundamentales del propio profesional.

7.

8. PROPUESTAS PARA ACTUALIZACIÓN CODIGO DEONTOLÓGICO.

El Código de Deontología Médica de 2022 ya incorpora referencias relevantes a la tecnología, la telemedicina y la inteligencia artificial (arts. 81–86, especialmente el capítulo XXIV "Inteligencia artificial (IA) y bases de datos sanitarios"). Sin embargo, la aparición del AI Act (Reglamento [UE] 2024/1689), que identifica explícitamente a los médicos y centros sanitarios como deployers de sistemas de alto riesgo, hace aconsejable un ajuste fino del Código para proteger de forma simétrica los derechos y obligaciones del paciente **y** del profesional.

A partir de los desarrollos previos de este capítulo, pueden proponerse los siguientes ejes de reforma:

1. Autonomía clínica y derecho a apartarse del algoritmo (Arts. 39, 42, 85 y 86):

El Código ya sitúa la seguridad del paciente como prioridad (art. 39) y exige encuadrar la práctica clínica en guías y protocolos, permitiendo apartarse de ellos cuando el caso lo requiera y dejando constancia en la historia clínica (art. 42.1–2). Los arts. 85–86 recuerdan que la IA y las grandes bases de datos son herramientas de ayuda y no sustituyen la obligación de buena práctica profesional.

No obstante, el Código podría explicitar con mayor claridad que, frente a sistemas de apoyo a la decisión basados en IA, la decisión clínica final es siempre humana y que el médico **tiene derecho y deber** de apartarse razonadamente de la recomendación algorítmica cuando lo exija la lex artis.

Propuesta de modificación:

- Añadir un nuevo apartado al art. 42 (p. ej. 42.3):

«En el uso de sistemas de ayuda a la decisión clínica, incluidos los basados en inteligencia artificial, el médico mantendrá en todo momento su autonomía profesional para aceptar, matizar o rechazar las recomendaciones emitidas por dichos sistemas. Cuando se aparte de ellas por considerar que no se ajustan a la situación concreta del paciente o a la mejor evidencia disponible, deberá dejar constancia razonada en la historia clínica.»

- Matizar el art. 86 con un nuevo inciso (p. ej. 86.6):

«Los sistemas de inteligencia artificial y los datos procedentes de grandes bases de datos sanitarias no podrán ser utilizados para menoscabar la autonomía clínica del médico ni para imponer decisiones automatizadas contrarias a su juicio profesional responsable.»

Con ello, el Código se alinea con el AI Act, que exige que los sistemas de alto riesgo estén sometidos a supervisión humana efectiva, incluyendo la posibilidad de intervención y parada.

2. Lex artis digital y condición de deployer (Arts. 23, 24, 39, 56–57 y 85–86):

El Código ya sanciona el uso de prácticas carentes de base científica y pseudoterapias (art. 23), y exige conducta íntegra, diligente y competente (art. 24.1). Sin embargo, no explicita qué se espera de un médico o institución que actúa como deployer de sistemas de IA, figura ahora central en el AI Act.

Propuesta de modificación:

- Nuevo apartado en art. 23 (p. ej. 23.4):

«La lex artis incluye el uso prudente y crítico de herramientas digitales y sistemas de inteligencia artificial que cuenten con la debida validación científica, regulatoria y ética. El médico debe abstenerse de utilizar sistemas de ayuda a la decisión clínica que carezcan de evidencia suficiente, no se ajusten al contexto

asistencial en el que se aplican o no ofrezcan garantías adecuadas de seguridad y transparencia.»

- **Nuevo apartado en art. 56 (p. ej. 56.3), dirigido a instituciones y directivos:**

«Las instituciones sanitarias que implanten sistemas de inteligencia artificial o automatización clínica actúan como responsables de su despliegue y deberán garantizar su evaluación previa, la gestión continuada de riesgos, la existencia de procedimientos de supervisión humana y la posibilidad de suspender su uso cuando exista sospecha razonable de riesgo para los pacientes o para los profesionales.»

- **Reforzar el art. 85 con una referencia expresa al papel del médico como desplegador (deployer):**

«Cuando el médico o la institución sanitaria utilicen sistemas de inteligencia artificial, deberán cumplir las obligaciones inherentes a su condición de usuarios responsables de sistemas de alto riesgo, de acuerdo con la normativa europea vigente, incluyendo la evaluación de impacto, la formación de los profesionales y la trazabilidad de su uso.»

3. Condiciones de trabajo seguras en entornos altamente digitalizados (Arts. 39–43 y 56–57):

La obligación de priorizar la seguridad del paciente (art. 39) y de notificar incidentes y eventos adversos (arts. 41–43) está bien desarrollada. Los arts. 56–57 imponen a las instituciones la creación de condiciones adecuadas de calidad, seguridad y suficiencia.
En la práctica, la introducción precipitada de sistemas de IA puede generar nuevas formas de riesgo: alert fatigue, dependencia excesiva de recomendaciones automatizadas, infra-dotación de recursos humanos basada en expectativas irreales de "eficiencia algorítmica", etc.

Propuesta de modificación:

- Añadir un inciso a art. 39:

«La introducción de nuevas tecnologías y sistemas de inteligencia artificial no debe comprometer la seguridad del paciente ni las condiciones de

trabajo del médico. Es contrario a la Deontología que se implanten sistemas tecnológicamente sofisticados sin la adecuada validación, dotación de recursos y soporte técnico.»

- Nuevo apartado en art. 41 (p. ej. 41.4):

«Los incidentes y eventos adversos vinculados al uso de sistemas de inteligencia artificial, robótica o automatización clínica deberán ser notificados y analizados de forma específica, con el fin de corregir fallos sistémicos y mejorar su diseño e implementación.»

- Nuevo apartado en art. 56 (p. ej. 56.4):

«Es contrario a la Deontología exigir al médico el uso de sistemas de inteligencia artificial sin soporte técnico suficiente, sin mecanismos claros de parada o suspensión, o cuando estos generen una carga de trabajo desproporcionada o insegura.»
Estas modificaciones conectan el deber de seguridad del paciente con la obligación de ofrecer al médico un entorno de trabajo seguro también frente a riesgos digitales emergentes, en coherencia con la normativa europea sobre prevención de riesgos laborales y con las recomendaciones de la OMS sobre gobernanza de la IA en salud.

4. Formación continuada y alfabetización en IA (Arts. 77–78 y 85–86):

El Código reconoce la formación continuada como deber y derecho del médico (art. 77.1) y subraya la importancia de la ética y la deontología en la docencia (arts. 77–78). Sin embargo, no menciona de forma expresa la alfabetización digital y en IA, pese a que el AI Act incluye la capacitación de los deployers como elemento clave de supervisión humana eficaz, y la OMS y la WMA recomiendan programas específicos de formación en IA para profesionales sanitarios.

Propuesta de modificación:

- Nuevo apartado en art. 77 (p. ej. 77.3):

«La formación médica continuada debe incluir de manera progresiva competencias en salud digital e inteligencia artificial, abarcando el

conocimiento básico de su funcionamiento, sus beneficios y riesgos, las implicaciones ético-legales y las habilidades necesarias para su uso crítico y responsable en la práctica clínica.»

- Nuevo apartado en art. 78 (p. ej. 78.5):

«Los programas de formación de grado y posgrado incorporarán contenidos específicos sobre ética de la digitalización sanitaria, uso de datos masivos e inteligencia artificial, con especial atención a la preservación de la autonomía clínica, la seguridad del paciente y la equidad.»

- Refuerzo del art. 85:

«Cuando el médico participe en proyectos de investigación o implementación de sistemas de inteligencia artificial, deberá haber recibido formación adecuada y actualizada sobre los aspectos técnicos y éticos relevantes.»

5. Objeción de conciencia tecnológica y objeción de ciencia (Arts. 34–37, 35.1, 36.3 y 76.5):

El Código regula la objeción de conciencia (arts. 34–37) y reconoce la "objeción de ciencia" para el médico perito que no se considera capacitado (art. 76.5). En el contexto de la IA, surge la necesidad de contemplar una objeción específica frente a sistemas que el médico considera inseguros, sesgados o insuficientemente validados, sin que ello equivalga a un rechazo global de la tecnología.

Propuesta de modificación:

- Añadir un nuevo apartado al art. 35 (p. ej. 35.3):

«El médico podrá ejercer una objeción de conciencia profesional frente al uso de sistemas de inteligencia artificial u otras tecnologías automatizadas cuando, de manera razonada y documentada, considere que su utilización compromete la seguridad, la dignidad o la adecuada atención del paciente, o cuando no existan garantías suficientes de validez científica, transparencia o supervisión humana.»

- Nuevo inciso en art. 36:

«El médico que invoque objeción de conciencia tecnológica deberá garantizar que el paciente recibe atención alternativa adecuada y no será objeto de represalias, siempre que su objeción se fundamente en criterios de buena práctica clínica y en la protección de los derechos del paciente.»

- Extender la referencia a la objeción de ciencia del art. 76.5 al contexto tecnológico:

«La objeción de ciencia incluye la negativa razonada a utilizar sistemas diagnósticos o terapéuticos, incluidos los basados en inteligencia artificial, para los que el médico considere que no existe capacitación suficiente, evidencia robusta o condiciones de uso seguro.»

Con ello, el Código se armoniza con posiciones emergentes en otros ordenamientos que reconocen el derecho del profesional a rechazar sistemas de IA opacos o inseguros, siempre que lo haga en beneficio del paciente y no de forma arbitraria.

6. Logs, métricas de desempeño y protección de los datos del profesional (Arts. 27–33, 39–43 y 85–86):

El Código protege de forma sólida la confidencialidad de los datos del paciente (cap. VII), pero no contempla de manera explícita el estatuto ético de los datos generados por los propios sistemas de IA sobre la actuación del médico (logs de uso, métricas de rendimiento, índices de discrepancia, etc.). El AI Act, el RGPD y la LOPDGDD limitan las decisiones automatizadas en el ámbito laboral y exigen transparencia cuando se usan sistemas de alto riesgo para la gestión de personas trabajadoras.

Propuesta de modificación:

- Nuevo artículo o apartado dentro del capítulo VII (p. ej. art. 33.5):

«Los datos de registro (logs), métricas de desempeño y demás trazas digitales generadas por los sistemas de información sanitaria y de inteligencia artificial respecto a la actividad del médico deberán utilizarse primordialmente con fines de calidad asistencial, seguridad del paciente, investigación y mejora de los sistemas. Su utilización con fines disciplinarios, de control de productividad o de evaluación laboral deberá estar sujeta a criterios de

proporcionalidad, transparencia, información previa y respeto a la autonomía clínica, de acuerdo con la normativa de protección de datos aplicable.»

- Vincular el art. 41 sobre seguridad del paciente con esta garantía, añadiendo un inciso:

«La notificación y análisis de incidentes vinculados a la tecnología no legitimará prácticas de vigilancia laboral que menoscaben la confianza profesional ni la libertad de criterio del médico.»

7. Responsabilidad proporcional y seguro de responsabilidad civil en ecosistemas de IA

(Art. 24.4 y capítulo VI):

El Código exige que el médico disponga de seguro de responsabilidad civil profesional (art. 24.4), pero fue elaborado antes de la aprobación de la nueva Directiva de responsabilidad por productos defectuosos (Directiva [UE] 2024/2853), que moderniza la responsabilidad objetiva para abarcar software y sistemas de IA, y antes de la plena configuración del AI Act.
En un entorno en el que concurren múltiples actores (fabricantes de software, proveedores de datos, hospitales como deployers, etc.), el Código podría aclarar que la imputación de responsabilidad debe ser proporcional al rol desempeñado por cada agente.

Propuesta de modificación:

- Nuevo apartado en art. 24 (p. ej. 24.5):

«En los supuestos en que un daño al paciente esté relacionado con el uso de dispositivos, programas o sistemas de inteligencia artificial, la responsabilidad profesional del médico deberá valorarse teniendo en cuenta su actuación conforme a la lex artis, el grado de control efectivo sobre el sistema y el cumplimiento de sus deberes de supervisión y notificación. No es deontológicamente aceptable atribuir al médico, de forma automática y exclusiva, daños derivados principalmente de defectos de diseño, implementación o mantenimiento de dichos sistemas.»

- Nuevo apartado en art. 24 (p. ej. 24.6):

«Los Colegios y las instituciones sanitarias promoverán que los seguros de responsabilidad civil profesional y las pólizas institucionales contemplen de forma adecuada los riesgos específicos asociados al uso de tecnologías digitales e inteligencia artificial, de modo que el médico no quede desproporcionadamente expuesto frente a fallos sistémicos ajenos a su control.»

8. Libertad académica, denuncia de riesgos y canales de alerta sobre IA (Arts. 41–43, 77–79):

El Código ya obliga a comunicar incidentes y eventos adversos y a mejorar la seguridad del paciente (arts. 41–43) y regula las publicaciones profesionales (art. 79). La Directiva (UE) 2019/1937 y la Ley 2/2023 protegen a quienes informan sobre infracciones normativas, creando canales internos y externos de información sin represalias.
En el contexto de la IA, es crucial que los médicos puedan comunicar sesgos, fallos o riesgos sistemáticos de los sistemas sin temor a consecuencias laborales o contractuales.

Propuesta de modificación:

- Nuevo apartado en art. 41 (p. ej. 41.5):

«El médico que detecte riesgos relevantes para la seguridad del paciente derivados del uso de sistemas de inteligencia artificial u otras tecnologías digitales debe poder comunicarlos, a través de los canales internos o externos previstos en la normativa, sin sufrir represalias laborales o profesionales, de acuerdo con las leyes de protección de informantes y la normativa de dispositivos sanitarios.»

- Nuevo apartado en art. 79 (p. ej. 79.4):

«Es conforme a la Deontología Médica que el médico publique resultados, series de casos o análisis críticos sobre el funcionamiento, limitaciones o riesgos de sistemas de inteligencia artificial y otras tecnologías sanitarias, siempre que respete la confidencialidad de los pacientes, comunique previamente los riesgos a las autoridade competentes cuando sea exigible y evite un tratamiento sensacionalista de la información.»
Estas garantías refuerzan la función del médico como “sensor ético” del sistema sanitario y como actor clave en la vigilancia poscomercialización de las herramientas de IA.

9. BIBLIOGRAFÍA

1. Trujillo Ruiz JA. Guía básica de Inteligencia Médica Artificial. Independently published; 2023. ISBN: 9798862907230. Amazon España+1

2. Trujillo Ruiz JA. Inteligencia Artificial y Derechos de los Pacientes: el equilibrio necesario. Independently published; 2025. ISBN: 9798313198897. Amazon+1

3. Risse M. The Fourth Generation of Human Rights: Epistemic Rights in the Digital Age. Harvard Kennedy School; 2021.appext.hks.harvard.edu

4. Razmetaeva Y. The concept of human rights in the digital era and bio-information rights as fourth generation rights. Asian J Legal Ethics. 2022.ajee-journal.com

5. International IDEA. Rights in the Digital Age. Estocolmo: International IDEA; 2025.idea.int

6. Van Kolfschooten HB, et al. A health-conformant reading of the GDPR's right not to be subject to automated decision-making. Health Policy Technol. 2024.PMC

7. Funer F. Physician's autonomy in the face of AI support. BMC Med Ethics. 2024.PMC

8. Jones C, et al. Artificial intelligence and clinical decision support: clinicians' responsibilities in the age of AI. Med Law Rev. 2023;31(4):501-524.OUP Academic

9. Größer J, et al. Studying the potential effects of artificial intelligence on physician autonomy and decision-making. Healthcare Analytics. 2025.ScienceDirect

10. Šustek P, Šolc M. Civil liability for artificial intelligence in medicine: is there a need for a new paradigm? The Lawyer Quarterly. 2025;34(3):464-484.tlq.ilaw.cas.cz

11. Giorgetti C, et al. Establishing new boundaries for medical liability: the role of AI-based clinical decision support under the EU AI Act. Adv Clin Exp Med. 2025;34(10):1601-1611.advances.umw.edu.pl

12. Van Leeuwen KG, et al. The AI Act: responsibilities and obligations for healthcare professionals and organizations. Diagn Interv Radiol. 2025.dirjournal.orgEuropean Commission. Artificial Intelligence in healthcare. Dirección General de Salud y Seguridad Alimentaria; 2024.Public Health

13. World Medical Association. Ethics, Legal, and Regulatory Aspects of AI in Healthcare – Summary Document. WMA; 2025.wma.net

14. Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (AI Act). Texto en EUR-Lex (versión en español): https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=CELEX:32024R1689

15. Reglamento (UE) 2025/327 del Parlamento Europeo y del Consejo, de 11 de febrero de 2025, relativo al Espacio Europeo de Datos de Salud (EEDS).
Texto en EUR-Lex (versión en español):
https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=CELEX:32025R0327 EUR-Lex

16. Van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): Implications for healthcare. Health Policy. 2024;149:105152.
Ficha y acceso vía repositorio UvA-DARE:
https://dare.uva.nl/personal/pure/en/publications/the-eu-artificial-intelligence-act-2024-implications-for-healthcare(e4cce4ba-ec0e-4fbd-902f-e414ad1dc056).html dare.uva.nl

17. Van Leeuwen KG, Doorn L, Gelderblom E. The AI Act: responsibilities and obligations for healthcare professionals and organizations. Diagn Interv Radiol. 2025 (online first). PubMed: https://pubmed.ncbi.nlm.nih.gov/40439140/ PubMed

18. Busch F, Kather JN, Johner C, et al. Navigating the European Union Artificial Intelligence Act for Healthcare. npj Digital Medicine. 2024;7(1):116.
Versión en npj Digital Medicine (Nature):
https://www.nature.com/articles/s41746-024-01213-6 Nature

19. Busch F, et al. AI policy in healthcare: a checklist-based methodology for hospital governance under the EU AI Act. Nature

20. Ministerio de Sanidad (España). Consulta pública previa sobre el Anteproyecto de Ley de Salud Digital. Secretaría General de Salud Digital, Información e Innovación del SNS.
Portal del Ministerio / participación pública (los enlaces concretos pueden cambiar, pero se accede a través de "Participación pública – Anteproyecto de Ley de Salud Digital"):
https://www.sanidad.gob.es/ (buscar "Ley de Salud Digital" en el apartado de participación pública o normativa en tramitación). BOE

21. European Commission. Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI), COM(2025) 836 final. Brussels, 19 Nov 2025. Disponible en:
https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A52025PC0836

22. European Commission. Simpler EU digital rules and new digital wallets to save billions for businesses and boost innovation (Digital Package press release, IP_25_2718). Press corner, 19 Nov 2025. Disponible en: https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2718 Estrategia Digital Europea

23. European Commission – DG CONNECT. Digital Package – Frequently Asked Questions. Shaping Europe's Digital Future, 20 Nov 2025. Disponible en: https://digital-strategy.ec.europa.eu/en/faqs/digital-package Estrategia Digital Europea

24. European Commission – DG CONNECT. Digital Omnibus Regulation Proposal – Proposal for Regulation on simplification of the digital legislation (Digital Package on Simplification). Shaping Europe's Digital Future, 19 Nov 2025. Disponible en: https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal Estrategia Digital Europea

25. Simmons & Simmons. AI View: November 2025 – European Commission proposes amendments to EU AI Act (Digital Omnibus update). 26 Nov 2025. Disponible en: https://www.simmons-simmons.com/en/publications/cmig1vhxv004oujfota7u217x/ai-view-november-2025Simmons & Simmons

26. White & Case LLP. EU Digital Omnibus: What changes lie ahead for the Data Act, GDPR and AI Act. 21 Nov 2025. Disponible en: https://www.whitecase.com/insight-alert/eu-digital-omnibus-what-changes-lie-ahead-data-act-gdpr-and-ai-actwhitecase.com

27. Clifford Chance. All aboard the Digital Omnibus? An overview of the EU's Digital Simplification Package. 25 Nov 2025. Disponible en: https://www.cliffordchance.com/briefings/2025/11/overview-of-the-eu-digital-simplification-package.htmlClifford Chance

28. Morrison & Foerster LLP. EU Digital Omnibus on AI: What Is in It and What Is Not? 2025. Disponible en: https://www.mofo.com/resources/insights/251201-eu-digital-omnibus Morrison Foerster

29. Mukherjee S, Meijer BH. EU to delay “high risk” AI rules until 2027 after Big Tech pushback. Reuters, 19 Nov 2025. Disponible en: https://www.reuters.com/sustainability/boards-policy-regulation/eu-delay-high-risk-ai-rules-until-2027-after-big-tech-pushback-2025-11-19/ Reuters

30. de Falguera D. El dilema del “digital omnibus”: agilidad frente a seguridad jurídica. Legal – Cinco Días (El País), 5 Dic 2025. Disponible en: https://cincodias.elpais.com/legal/2025-12-05/el-dilema-del-digital-omnibus-agilidad-frente-a-seguridad-juridica.html Cinco Días

31. European Digital Rights (EDRi). Why the Digital Omnibus puts GDPR and ePrivacy at risk. 2025. Disponible en: https://edri.org/our-work/why-the-digital-omnibus-puts-gdpr-and-eprivacy-at-risk/ European Digital Rights (EDRi)

32. Schmon C, Gullo K. EU’s New Digital Package Proposal Promises Red Tape Cuts but Guts GDPR Privacy Rights.Electronic Frontier Foundation (Deeplinks Blog), 4 Dec 2025. Disponible en: https://www.eff.org/deeplinks/2025/12/eus-new-digital-package-proposal-promises-red-tape-cuts-guts-gdpr-privacy-rights Electronic Frontier Foundation

33. Unión Europea. Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (AI Act). DOUE L 2024/1689, 12.7.2024.EUR-Lex+1

34. Artificial Intelligence Act website. “Article 3 – Definitions” y “Article 26 – Obligations of deployers of high-risk AI systems”. artificialintelligenceact.eu+1

35. García Luengo J. “Obligations of providers and deployers of high-risk AI systems (Chapter III, Section 3)”. En: Obligations of providers and deployers of high-risk AI systems. Wolters Kluwer; 2025.Dialnet

36. Van Leeuwen KG, Doorn L, Gelderblom E. “The AI Act: responsibilities and obligations for healthcare professionals and organizations”. Diagnostic and Interventional Radiology. 2025. dirjournal.org+1

37. Bignami E et al. “AI policy in healthcare: a checklist-based methodology for implementing the EU AI Act in high-acuity settings”. 2025.PMC

38. Osborne Clarke. “EU AI Act’s ‘deployers’ definition has wide-ranging significance for life sciences”. 2024.osborneclarke.com

39. World Medical Association. WMA International Code of Medical Ethics. 2022.

40. World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. 2024–2025.

41. Wong A et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model (Epic Sepsis Model). JAMA Intern Med. 2021.

42. Habib AR. The Epic Sepsis Model Falls Short—The Importance of External Validation. JAMA Intern Med. 2021. https://doi.org/10.1001/jamainternmed.2021.3333

43. Ross C, Swetlitz I. IBM’s Watson Recommended ‘Unsafe and Incorrect’ Cancer Treatments. STAT. 2018.

44. Fink M. Human Oversight under Article 14 of the EU AI Act. In: Malgieri G, González Fuster G, Mantelero A, Zanfir-Fortuna G (eds.). AI Act Commentary: A Thematic Analysis (Hart, en prensa).SSRN

45. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act).EurLex+2streamlex.eu+2

46. Funer F, Wiesing U. Physician's autonomy in the face of AI support: walking the ethical tightrope. Frontiers in Medicine. 2024;11:1324963. https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2024.1324963/full

47. Council of Europe. Convention on Human Rights and Biomedicine (Oviedo Convention), 1997.Wikipedia+1

48. Organización Médica Colegial de España. Código de Deontología Médica: Guía de Ética Médica. CGCOM, 2011 y actualización 2022.comteruel.es+1

49. Díez JA. El nuevo código deontológico se amplía con una objeción que deja varias sombras. Diario Médico, 2011.https://simeg.org

50. Herreros B, et al. Guidelines for conscientious objection in Spain: a proposal for healthcare professionals. BMC Med Ethics. 2024;25:22.PMC

51. Wicclair M. Conscientious Objection in Healthcare and Moral Integrity. Cambridge Q Healthc Ethics. 2017;26(1):23-36. Cambridge University Press & Assessment

52. Celie KB, et al. Conscientious objection: a global health perspective. BMJ Global Health. 2024;9:e017555.gh.bmj.com

53. World Medical Association. WMA Statement on Artificial and Augmented Intelligence in Medical Care, revisada 2024/2025. Asociación Médica Mundial

54. American Medical Association. Augmented Intelligence in Health Care (Policy H-480.940).American Medical Association

55. Haltaufderheide J, Ranisch R. The Ethics of ChatGPT in Medicine and Healthcare: A Systematic Review on Large Language Models. 2024.arXiv

56. CPSO (College of Physicians and Surgeons of Ontario). Guidance on the use of AI in clinical practice, 2024.

57. Directiva (UE) 2024/2853 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, sobre responsabilidad por los daños causados por productos defectuosos.Eur-Lex+1

58. Organización Mundial de la Salud. Ethics and Governance of Artificial Intelligence for Health. Ginebra: OMS; 2021. Organización Mundial de la Salud+1

59. Organización Mundial de la Salud. Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models. Ginebra: OMS; 2025.Organización Mundial de la Salud

60. World Medical Association. International Code of Medical Ethics. WMA; 2022.wma.net+1

61. World Medical Association. Statement on Artificial and Augmented Intelligence in Medical Care. WMA; 2025.wma.ne

8. ANEXOS:

1. Contrato tipo entre el médico y su institución sanitaria en la que trabaja.

ANEXO CONTRACTUAL (Propuesta):

Derechos del profesional médico y condiciones de uso de sistemas de IA clínica en una institución sanitaria.

Entre

- [EL HOSPITAL/LA ENTIDAD SANITARIA], con domicilio en [•••], CIF [•••], representada por [•••], en adelante, EL HOSPITAL;
- [EL PROVEEDOR TECNOLÓGICO], con domicilio en [•••], CIF [•••], representada por [•••], en adelante, EL PROVEEDOR;
- Y, a los solos efectos de reconocimiento expreso de derechos como tercero beneficiario, [EL/LA PROFESIONAL MÉDICO/A], colegiado/a nº [•••], en adelante, EL PROFESIONAL.

Objeto. Establecer las condiciones contractuales para la adquisición, despliegue y uso de sistemas de inteligencia artificial (IA) en la práctica clínica, garantizando los derechos del profesional médico y el cumplimiento del marco legal aplicable (Reglamento de IA de la UE, normativa de productos sanitarios, RGPD/LOPDGDD, PRL y demás).

7 Derechos y obligaciones del médico ante la IA

Dr. José Antonio Trujillo

1. Definiciones operativas

- Sistema de IA: herramienta software con funciones de predicción, clasificación, recomendación o generación que influye en decisiones clínicas.
- Alto riesgo: los sistemas destinados a apoyar decisiones diagnósticas/terapéuticas u organizar la atención con impacto significativo.
- Supervisión humana / override: capacidad del profesional de ignorar, desactivar o revertir la salida de IA sin penalización.
- Logs: registros técnicos y clínicos necesarios para trazabilidad, seguridad y auditoría.
- Cambio sustancial: modificación con impacto en desempeño, indicaciones, datos, explicabilidad, interfaz o integración.

2. Principio de autonomía clínica y supervisión humana

2.1. EL PROVEEDOR garantiza botón de parada y controles de override en todos los flujos donde la IA pueda influir en decisiones asistenciales.

2.2. EL PROFESIONAL podrá ignorar o desactivar la salida de la IA por razones clínicas sin menoscabo económico, disciplinario ni reputacional.

2.3. EL HOSPITAL O INSTITUCÍON SANITARIA adoptará una política escrita: la IA informa; no decide, incorporándola a protocolos y formación.

2.4. Toda salida de IA incluirá explicabilidad básica (variables relevantes, justificación, confianza y límites de uso) accesible en la historia clínica.

3. Lex artis, validación local y gestión de cambios

3.1. Antes del uso clínico, EL HOSPITAL O INSTITUCIÓN SANIATARIA realizará validación local (población, entorno técnico, integración) con métricas verificables.

3.2. EL PROVEEDOR entregará documento de uso previsto, contraindicaciones, model card (datos, desempeño, sesgos, ciberseguridad) y manual de riesgos.

3.3. Todo cambio sustancial exigirá pruebas controladas, aprobación del Comité de IA y trazabilidad por versión.

3.4. EL PROFESIONAL conserva la última palabra; la IA no desplaza la lex artis ni la responsabilidad clínica individual dentro de sus límites.

4. Condiciones de trabajo seguras y mantenimiento (SLA)

4.1. EL PROVEEDOR mantendrá un SLA mínimo de disponibilidad [≥ 99,X%] y MTTR ≤ [Y] horas, con parches de seguridad críticos ≤ [Z] días.

4.2. Existirá registro de riesgos de IA, monitorización de drift, canal de incidentes y plan de contención y rollback.

4.3. EL HOSPITAL O INSTITUCIÓN SANITARIA realizará evaluación de PRL por introducción de IA (carga cognitiva, fatiga de alarmas, ergonomía y factores humanos) y adoptará medidas.

4.4. EL PROVEEDOR otorgará acceso a logs auditables (protegiendo datos personales) y a informes de vigilancia posdespliegue.

5. Formación y acreditación previa

5.1. El acceso de EL PROFESIONAL a la IA exigirá formación pagada y acreditada (fundamentos, sesgos, ciber, DPIA, override, límites de uso).

5.2. Las actualizaciones con impacto clínico requerirán formación de refresco previa a su activación.

5.3. EL HOSPITAL O INSTITUCIÓN SANITARIA mantendrá registro de acreditaciones vigentes y simulaciones de fallo/contingencia.

6. Responsabilidad proporcional e indemnidad

6.1. EL PROVEEDOR asume la responsabilidad e indemnidad por defectos del producto software (diseño, desarrollo, actualización, ciberseguridad), con seguro suficiente

6.2. EL HOSPITAL O INSTITUCIÓN SANITARIA responde por uso indebido o por incumplir validación, monitorización, formación o procedimientos aprobados.

6.3. Las partes cooperarán en investigaciones y peritajes, con acceso a artefactos de auditoría, preservación de evidencia y cadena de custodia.

6.4. No se trasladarán a EL PROFESIONAL fallos atribuibles al producto ni carencias documentadas del sistema.

7. Libertad académica, seguridad del paciente y alertas

7.1. EL HOSPITAL O INSTITUCIÓN SANITARIA establecerá canal protegido de alertas (whistleblowing) y comité técnicoclínico para revisión y acción.

7.2. EL PROFESIONAL podrá notificar fallos, sesgos o riesgos sin represalias; toda denuncia será tramitada con confidencialidad y plazos.

7.3. Tras la notificación y ventana razonable de mitigación, EL PROFESIONAL tendrá derecho a divulgar hallazgos no personales por interés científico y de seguridad.

8. Protección de datos del profesional y uso de logs

8.1. Se realizará DPIA específica; los logs se limitarán a lo necesario para seguridad y calidad. Queda prohibido su uso como vigilancia laboral o para decisiones automatizadas con efectos profesionales.

8.2. Transparencia: se informará qué se registra, quién accede, para qué y plazos de retención [≤ ••• meses].

8.3. Los indicadores de desempeño se anonimizarán o agregarán; cualquier evaluación individual tendrá revisión humana y derecho de impugnación.

9. Gobernanza, auditoría y comité de IA

9.1. Se crea un Comité de IA (clínico, TIC, legal, calidad, seguridad del paciente) que aprueba despliegues, cambios y planes de mitigación.

9.2. EL PROVEEDOR permitirá auditorías razonables sobre seguridad, desempeño y cumplimiento, preservando secretos industriales.

9.3. Se adoptarán marcos de gestión de riesgo (p. ej., NIST AI RMF / ISO 42001) y un cuadro de mando con KPIs.

10. Ciberseguridad y continuidad

10.1. Política de divulgación responsable de vulnerabilidades y canal técnico 24×7. 10.2. Plan de continuidad y procedimiento de degradación segura ante caída del sistema, garantizando la atención.

11. Propiedad intelectual y datos

11.1. La PI del sistema y pesos de modelos corresponde a EL PROVEEDOR; las historias clínicas y datos pertenecen a EL HOSPITAL O INSTITUCIÓN SANITARIA y/o a los pacientes según ley.

11.2. Queda prohibido entrenar o reentrenar con datos personales sin base legal, consentimiento cuando proceda y medidas de anonimización/pseudonimización.

11.3. Salvo pacto expreso, no se conceden licencias implícitas sobre datasets clínicos.

12. Confidencialidad

12.1. Las partes guardarán confidencialidad sobre información técnica, clínica y de seguridad, sin perjuicio de obligaciones regulatorias de reporte.

13. Suspensión, terminación y plan de salida

13.1. EL HOSPITAL O INSTITUCIÓN SANITARIA podrá suspender el uso ante riesgos graves, drift no controlado o incumplimientos de SLA/ciberseguridad.

13.2. A la terminación: plan de salida con portabilidad de configuraciones, exportación de logs y documentación, y asistencia razonable durante [•••] días.

14. Cumplimiento normativo, ley aplicable y jurisdicción

14.1. Las partes cumplirán la normativa aplicable (Reglamento de IA de la UE, MDR/IVDR cuando proceda, RGPD/LOPDGDD, PRL, derecho sanitario y deontología).

14.2. Ley aplicable: Derecho español. Jurisdicción: tribunales de [•••], sin perjuicio de foros imperativos.

Firmas:

En [•••], a [•••] de [•••] de [20••].
EL HOSPITAL O INSTITUCIÓN SANITARIA
Nombre: _____
Cargo: _____
Firma: _____
EL PROVEEDOR
Nombre: _____
Cargo: _____
Firma: _____
EL/LA PROFESIONAL (tercero beneficiario)
Nombre: _____
Nº colegiado: _____
Firma: _____

15. Check list derechos fundamentales del médico con respecto a la ia

Checklist operativo — Derechos del médico en la IA clínica

1) Autonomía clínica:

Requisitos mínimos

- "Botón de parada" y override documentado en todos los sistemas de IA.
- Explicabilidad básica accesible en la historia (por qué, con qué datos, confianza).

- Política escrita: la IA informa; no decide.

KPIs

- % de recomendaciones ignoradas/ajustadas con justificación clínica.
- Tiempo medio de respuesta al override crítico.
- Incidentes evitados por decisión clínica frente a IA.

Cláusulas contractuales

- "El usuario clínico podrá desactivar/ignorar la salida sin penalización."
- "El proveedor garantiza controles de supervisión humana y registro del override."
- "La IA no constituirá mandato clínico ni sustituirá el juicio profesional."

2) **Debida diligencia (lex artis) y última palabra:**

Requisitos mínimos

- Validación local (desempeño, poblaciones, entorno técnico) antes del despliegue.
- Documento de uso previsto, límites y contraindicaciones clínicas.
- Gobierno de cambios: pruebas, aprobación clínica y trazabilidad por versión.

KPIs

- Sensibilidad/especificidad/AUC vs. estándar y falsas alarmas por 1.000 pacientes.
- % decisiones clínicas documentadas como contrarias a la IA con resultado seguro.
- Tasa de drift detectado y tiempo hasta mitigación.

Cláusulas contractuales

- "El profesional conserva la última palabra; la IA es apoyo."
- "El proveedor no impondrá prácticas automatizadas sin validación clínica."
- "Se entregarán métricas verificables y límites de uso; su incumplimiento activa garantía."

3) **Condiciones de trabajo seguras:**

Requisitos mínimos

- Registro de riesgos de IA, monitoreo continuo y plan de mantenimiento (SLA).
- Logs técnicos y clínicos auditables; canal de incidentes y aprendizaje.
- Evaluación de PRL por introducción de IA (fatiga de alarmas, carga cognitiva).

KPIs

- Uptime y MTTR (tiempo de recuperación) del sistema.
- Tasa de incidentes y tiempo a contención/corrección.
- % versiones validadas antes de activación clínica.

Cláusulas contractuales

- "IA validada y mantenida; SLA mínimo X% y parches de seguridad en ≤Y días."
- "Acceso del hospital a logs y a informes de vigilancia posdespliegue."
- "Notificación previa de cambios sustanciales y derecho a suspender."

4) **Formación continua y alfabetización digital:**

Requisitos mínimos

- Capacitación pagada y acreditada antes del uso (fundamentos, sesgos, ciber, DPIA, override).
- Simulaciones/sandbox y manual de emergencias (fallos, caída del sistema).
- Reciclaje anual y formación específica por actualización mayor.

KPIs

- % de profesionales acreditados antes del acceso.
- Score de competencia postformación y a los 6–12 meses.
- Tiempo medio desde actualización → formación → autorización.

Cláusulas contractuales

- "El empleador financiará la formación acreditada y refrescos."
- "Las actualizaciones con impacto clínico requieren formación previa."
- "El acceso clínico queda condicionado a la acreditación vigente."

5) **Responsabilidad proporcional:**

Requisitos mínimos

- Matriz RACI (fabricante–hospital–clínico) y mapa de riesgos por fallo (producto, datos, uso).
- Seguro/indemnidad del proveedor por defectos de software y actualizaciones.
- Model card/documentación: datos, desempeño, sesgos, ciber, ciclo de vida.

KPIs

- % incidentes con causa raíz atribuible al proveedor vs. uso.
- Tiempo de resolución e indemnización en reclamaciones.
- Nº cambios sustanciales notificados con antelación.

Cláusulas contractuales

- "El proveedor indemnizará daños por defecto (diseño, actualización, ciber)."
- "Cooperación plena en peritajes y acceso a artefactos de auditoría."
- "Prohibido trasladar al profesional fallos del producto."

6) Libertad académica e intelectual (alertas de seguridad):

Requisitos mínimos

- Canal protegido para alertas (whistleblowing) y comité de revisión técnica-clínica.
- Procedimiento para publicar hallazgos de seguridad respetando confidencialidad.
- Política antirrepresalias y reconocimiento autoral.

KPIs

- Tiempo de respuesta a alertas y % con acción correctiva.
- Nº de hallazgos publicados o compartidos con reguladores.
- Cero reportes de represalia verificados.

Cláusulas contractuales

- "Queda garantizada la no represalia por alertas de seguridad."

- “Derecho a difundir evidencias de riesgo tras notificación y mitigación razonable.”
- “El proveedor facilitará datos no personales para reproducción de hallazgos.”

7) **Protección de datos del profesional:**

Requisitos mínimos

- DPIA específica; minimización y agregación de métricas de desempeño.
- Prohibición de decisiones automatizadas con efectos laborales (sanciones, incentivos).
- Transparencia: qué se registra, quién accede, retención limitada y control de acceso.

KPIs

- Nº accesos a logs por rol/mes y incidentes de privacidad.
- % métricas de desempeño anonimizadas/agrupadas.
- 0 decisiones automatizadas sobre empleo.

Cláusulas contractuales

- “Los logs se usarán para seguridad y calidad, no para control disciplinario.”
- “Derechos de acceso/impugnación y revisión humana de cualquier evaluación.”
- “Retención máxima X meses y borrado garantizado.

16. Tabla resumen:

1) **Autonomía clínica:**

Requisitos mínimos	**KPIs**	**Cláusulas tipo**
Botón de parada y override en todos los sistemas.	% recomendaciones ignoradas/ajustadas con justificación clínica.	"El usuario clínico podrá desactivar/ignorar la salida sin penalización."
Explicabilidad básica en historia clínica (razón, datos, confianza).	Tiempo medio de respuesta al override crítico.	"El proveedor garantiza supervisión humana y registro del override."
Política escrita: la IA informa, no decide.	Incidentes evitados por decisión clínica frente a IA.	"La IA no constituirá mandato clínico ni sustituirá el juicio profesional."

2) **Debida diligencia *(lex artis*) y última palabra:**

Requisitos mínimos	**KPIs**	**Cláusulas tipo**
Validación local (desempeño, población, entorno técnico).	Sensibilidad/especificidad/AUC vs. estándar y falsas alarmas/1.000 pacientes.	"El profesional conserva la última palabra; la IA es apoyo."
Documento de uso previsto, límites y contraindicaciones.	% decisiones clínicas documentadas como contrarias a la IA con resultado seguro.	"No se impondrán prácticas automatizadas sin validación clínica."

Gobierno de cambios: pruebas, aprobación clínica, trazabilidad por versión.	Tasa de drift detectado y tiempo hasta mitigación.	"Se entregarán métricas verificables y límites de uso; su incumplimiento activa garantía."

3) **Condiciones de trabajo seguras:**

Requisitos mínimos	**KPIs**	**Cláusulas tipo**
Registro de riesgos de IA, monitoreo continuo y plan de mantenimiento (SLA).	Uptime y MTTR del sistema.	"IA validada y mantenida; SLA mínimo X% y parches de seguridad en ≤Y días."
Logs técnicos y clínicos auditables; canal de incidentes y aprendizaje.	Tasa de incidentes y tiempo a contención/corrección.	"Acceso del hospital a logs e informes de vigilancia posdespliegue."
Evaluación de PRL por introducción de IA (fatiga de alarmas, carga cognitiva).	% versiones validadas antes de activación clínica.	"Notificación previa de cambios sustanciales y derecho a suspender."

4) **Formación continua y alfabetización digital:**

Requisitos mínimos	KPIs	Cláusulas tipo
Capacitación pagada y acreditada antes del uso (fundamentos, sesgos, ciber, DPIA, override).	% de profesionales acreditados antes del acceso.	"El empleador financiará la formación acreditada y refrescos."
Simulaciones/sandbox y manual de emergencias (fallos, caída del sistema).	Score de competencia postformación y a 6–12 meses.	"Actualizaciones con impacto clínico requieren formación previa."
Reciclaje anual y formación específica por actualización mayor.	Tiempo desde actualización → formación → autorización.	"El acceso clínico queda condicionado a la acreditación vigente."

5) **Responsabilidad proporcional:**

Requisitos mínimos	KPIs	Cláusulas tipo
Matriz RACI (fabricante–hospital–clínico) y mapa de riesgos (producto, datos, uso).	% incidentes con causa raízatribuible al proveedor vs. uso.	"El proveedor indemnizará daños por defecto (diseño, actualización, ciber)."
Seguro/indemnidad del proveedor por defectos de software y actualizaciones.	Tiempo de resolución e indemnización en reclamaciones.	"Cooperación plena en peritajes y acceso a artefactos de auditoría."
Model card/documentación: datos, desempeño, sesgos, ciber, ciclo de vida.	Nº cambios sustancialesnotificados con antelación.	"Prohibido trasladar al profesional fallos del producto."

6) **Libertad académica e intelectual (alertas de seguridad):**

Requisitos mínimos	**KPIs**	**Cláusulas tipo**
Canal protegido para alertas (whistleblowing) y comité de revisión técnico-clínica.	Tiempo de respuesta a alertasy % con acción correctiva.	"Queda garantizada la no represalia por alertas de seguridad."
Procedimiento para publicar hallazgos de seguridad respetando confidencialidad.	Nº de hallazgos publicados o compartidos con reguladores.	"Derecho a difundir evidencias de riesgo tras notificación y mitigación razonable."
Política antirrepresalias y reconocimiento autoral.	Cero reportes de represalia verificados.	"El proveedor facilitará datos no personales para reproducción de hallazgos."

7) **Protección de datos del profesional:**

Requisitos mínimos	**KPIs**	**Cláusulas tipo**
DPIA específica; minimización y agregaciónde métricas de desempeño.	Nº accesos a logs por rol/mes y incidentes de privacidad.	"Los logs se usarán para seguridad y calidad, no para control disciplinario."
Prohibición de decisiones automatizadas con efectos laborales (sanciones, incentivos).	% métricas de desempeño anonimizadas/agrupadas.	"Derechos de acceso/ impugnación y revisión humana de cualquier evaluación."
Transparencia: qué se registra, quién accede, retención limitada y control de acceso.	0 decisiones automatizadas sobre empleo.	"Retención máxima X meses y borradogarantizado."

8

Protección de datos personales y Sistemas de IA

Dra. Cristina Gil Membrado
Catedrática de Derecho Civil.
Universidad de las Islas Baleares

Resumen ejecutivo

La inteligencia artificial (IA) se ha convertido en una herramienta clave en el ámbito sanitario por su capacidad para analizar grandes volúmenes de datos clínicos y detectar patrones que apoyan el diagnóstico, el pronóstico o la toma de decisiones médicas. Sin embargo, su uso plantea importantes implicaciones en materia de protección de datos personales, especialmente porque la IA médica trata, en la mayoría de los casos, datos de salud, que el Reglamento General de Protección de Datos (RGPD) considera una categoría especial de datos sujeta a una protección reforzada.

Un sistema de IA médica basado en *machine learning* utiliza datos a lo largo de todo su ciclo de vida. En la fase de entrenamiento, el algoritmo aprende a partir de datos históricos de pacientes, como edad, diagnósticos, pruebas clínicas o imágenes médicas. Posteriormente, en la fase de validación, se emplean otros conjuntos de datos para comprobar si el modelo funciona correctamente y generaliza a contextos distintos. En la fase de prueba, se evalúa su rendimiento final antes del despliegue clínico. Finalmente, durante la fase de uso, el sistema procesa datos de entrada correspondientes a pacientes reales para generar recomendaciones o predicciones. En todas estas fases puede existir tratamiento de datos personales, lo que activa la aplicación del RGPD.

El RGPD exige que cualquier tratamiento de datos personales cuente con una base jurídica válida. En el ámbito sanitario, el tratamiento de datos de salud está prohibido por defecto, salvo que concurra alguna de las excepciones del artículo 9.2, como el consentimiento explícito del paciente, el interés público en el ámbito de la salud, la asistencia sanitaria o la investigación científica con garantías. Además, deben respetarse principios fundamentales como la limitación de la finalidad, la minimización de datos, la exactitud, la seguridad y la transparencia. Esto implica que un sistema de IA solo puede utilizar los datos estrictamente necesarios para la finalidad prevista y debe informar adecuadamente a las personas afectadas sobre el uso de sus datos.

Junto al RGPD, el Reglamento Europeo de Inteligencia Artificial (Ley de IA) introduce un marco específico basado en el riesgo. La IA médica suele calificarse como IA de alto riesgo, ya que puede afectar directamente a la salud y a la vida de las personas. Por ello, la Ley de IA impone obligaciones adicionales, como la gestión y mitigación de riesgos, el uso de datos de

entrenamiento de calidad y representativos, la documentación técnica, la transparencia, la supervisión humana efectiva y la garantía de exactitud, robustez y ciberseguridad. Mientras el RGPD protege los derechos de las personas como titulares de los datos, la Ley de IA se centra en que el sistema sea técnicamente seguro y fiable. Ambas normas se complementan.

Un elemento clave para conciliar innovación y protección de derechos es el Espacio Europeo de Datos de Salud (EHDS/EEDS). Este marco permite reutilizar datos de salud recogidos originalmente con fines asistenciales (uso primario) para otros fines en beneficio de la sociedad, como la investigación, la innovación o el entrenamiento de algoritmos de IA (uso secundario). El acceso a los datos no es libre: requiere la autorización de un organismo público competente y se realiza bajo condiciones estrictas.

En el uso secundario dentro del EEDS, los datos se someten a seudonimización, se eliminan identificadores directos y se limitan a las variables necesarias. El análisis y el entrenamiento de la IA se llevan a cabo en entornos seguros de procesamiento, donde los investigadores no pueden descargar los datos ni reidentificar a los pacientes. Solo se permite la salida de resultados agregados, métricas o informes, tras controles destinados a evitar riesgos de reidentificación.

Además, los pacientes conservan sus derechos en virtud del RGPD: acceso, rectificación, supresión, limitación u oposición, con los límites propios del ámbito sanitario. El diseño de los sistemas debe incorporar la protección de datos desde el diseño y por defecto, garantizando que la privacidad se tenga en cuenta desde las primeras fases del desarrollo de la IA.

En conjunto, la combinación del RGPD, la Ley de IA y el EEDS configura un modelo europeo en el que la IA médica puede desarrollarse y entrenarse con datos reales de salud, pero siempre bajo un enfoque de seguridad, control, supervisión humana y respeto a los derechos fundamentales, generando confianza tanto en los profesionales sanitarios como en los pacientes.

Palabras clave

Inteligencia artificial médica; Protección de datos personales; Datos de salud; Reglamento General de Protección de Datos (RGPD); Ley Europea de Inteligencia Artificial (Ley de IA); Espacio Europeo de Datos de Salud (EEDS / EHDS)

Executive summary

Artificial intelligence (AI) has become a key tool in the healthcare sector due to its ability to analyse large volumes of clinical data and identify patterns that support diagnosis, prognosis, and medical decision-making. However, its use raises significant implications for personal data protection, particularly because medical AI systems usually process health data, which the General Data Protection Regulation (GDPR) classifies as a special category of personal data subject to enhanced protection.

A medical AI system based on machine learning uses data throughout its entire lifecycle. In the training phase, the algorithm learns from historical patient data, such as age, diagnoses, clinical tests, or medical images. In the validation phase, other datasets are used to check whether the model works correctly and generalises to different contexts. In the testing phase, its final performance is assessed before clinical deployment. Finally, during the use or deployment phase, the system processes input data relating to real patients in order to generate predictions or recommendations. In all these phases, personal data may be processed, triggering the application of the GDPR.

The GDPR requires that any processing of personal data be based on a valid legal basis. In healthcare, the processing of health data is prohibited by default unless one of the exceptions set out in Article 9(2) applies, such as the explicit consent of the patient, reasons of public interest in the field of public health, the provision of healthcare, or scientific research subject to appropriate safeguards. In addition, core GDPR principles must be respected, including purpose limitation, data minimisation, accuracy, security, and transparency. This means that an AI system may only use data that are strictly necessary for its intended purpose and that individuals must be properly informed about how their data are used.

Alongside the GDPR, the European Artificial Intelligence Act (AI Act) introduces a specific regulatory framework based on risk. Medical AI systems are generally considered high-risk AI systems, as they may have a direct impact on people's health and lives. As a result, the AI Act imposes additional obligations, such as risk management and mitigation, the use of high-quality and representative training data, technical documentation and traceability, transparency, effective human oversight, and guarantees

of accuracy, robustness, and cybersecurity. While the GDPR focuses on protecting individuals as data subjects, the AI Act concentrates on ensuring that AI systems are technically safe and trustworthy; the two frameworks are therefore complementary.

A key instrument for reconciling innovation with the protection of fundamental rights is the European Health Data Space (EHDS). This framework enables the reuse of health data originally collected for healthcare purposes (primary use) for other purposes that benefit society, such as research, innovation, policymaking, or the training, testing, and evaluation of AI algorithms (secondary use). Access to data under the EHDS is not free or unrestricted: it requires authorisation from a competent public authority and is subject to strict conditions.

In the context of secondary use under the EHDS, health data are subject to pseudonymisation, direct identifiers are removed, and datasets are limited to the variables strictly necessary for the approved purpose. Analysis and AI training take place within secure processing environments, where users cannot download the data or reidentify patients. Only authorised outputs, such as aggregated results, performance metrics, or scientific reports, may leave the secure environment, and only after checks designed to prevent reidentification risks.

Importantly, individuals retain their data protection rights under the GDPR, including the rights of access, rectification, erasure, restriction, and objection, subject to the specific limitations that apply in the healthcare and research context. Furthermore, medical AI systems must be designed in accordance with the principle of data protection by design and by default, ensuring that privacy safeguards are embedded from the earliest stages of development and throughout the system's lifecycle.

Taken together, the GDPR, the AI Act, and the EHDS establish a distinctively European model for medical AI. This model allows AI systems to be developed and trained using real health data, while ensuring a high level of protection for personal data through legal bases, minimisation, secure environments, human oversight, and enforceable rights. The result is a framework that seeks to foster innovation in healthcare AI while maintaining trust, legal certainty, and effective protection of patients' fundamental rights.

Keywords

Medical artificial intelligence; Personal data protection; Health data; General Data Protection Regulation (GDPR); European Artificial Intelligence Act (AI Act); European Health Data Space (EHDS)

Ideas fuerza:

La IA médica depende del tratamiento de datos de salud, que son datos personales especialmente sensibles y requieren una protección reforzada a lo largo de todo el ciclo de vida del sistema.

El RGPD es plenamente aplicable a la IA sanitaria cuando se tratan datos personales, exigiendo base jurídica, minimización de datos, transparencia, seguridad y respeto a los derechos de los pacientes.

La Ley de IA complementa al RGPD, calificando la IA médica como de alto riesgo y estableciendo obligaciones técnicas y organizativas para garantizar sistemas seguros, fiables y supervisados por humanos.

El EEDS permite el uso secundario de datos clínicos, de forma controlada y segura, para investigación e innovación, incluido el entrenamiento y la evaluación de algoritmos de IA.

La protección de datos debe integrarse desde el diseño y por defecto, mediante seudonimización, entornos seguros de procesamiento, control de accesos y supervisión humana efectiva, con el fin de generar confianza y proteger los derechos fundamentales.

Key messages:

Medical AI depends on the processing of health data, which are especially sensitive personal data and require enhanced protection throughout the entire lifecycle of the system.

8 Protección de datos personales y Sistemas de IA

Dra. Cristina Gil Membrado

The GDPR is fully applicable to healthcare AI when personal data are processed, requiring a legal basis, data minimization, transparency, security, and respect for patients' rights.

The AI Act complements the GDPR by classifying medical AI as high-risk and establishing technical and organizational obligations to ensure safe, reliable systems subject to human oversight.

The EHDS enables the secondary use of clinical data, in a controlled and secure manner, for research and innovation, including the training and evaluation of AI algorithms.

Data protection must be integrated by design and by default, through pseudonymization, secure processing environments, access control, and effective human oversight, in order to build trust and protect fundamental rights.

SUMARIO

1. INTROITO

La IA es un programa informático que se nutre de datos que alimentan al algoritmo, que no es sino un conjunto de instrucciones para resolver un problema. Su irrupción en la sociedad la convierte en una de las tecnologías disruptivas con más posibilidades, pero también con más riesgos por el impacto en los datos de carácter personal, y todavía más si el escenario es el sanitario y los datos que se tratan son datos personales relativos a la salud. A su vez, para pocos ámbitos como el sanitario, la IA tiene un potencial mayor.

El Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (RGPD) es la normativa que regula el tratamiento de datos personales de ciudadanos de la UE. Su objetivo principal es reforzar la privacidad, garantizando que el manejo de información sea transparente, seguro y proporcionar a los usuarios mayor control sobre sus datos.

El RGPD contiene un buen número de habilitaciones a los Estados miembros, para regular determinadas materias en la medida en que sea necesario por razones de coherencia y comprensión. Así, no se excluye toda intervención del Derecho interno en los ámbitos concernidos por los reglamentos europeos. Al contrario, tal intervención puede ser procedente, incluso necesaria, tanto para la depuración del ordenamiento nacional como para el desarrollo o complemento del reglamento de que se trate.

En este sentido, en España hay que mencionar la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales que tiene por objeto la adaptación del ordenamiento jurídico español al RGPD y la regulación del derecho fundamental de las personas físicas a la protección de datos personales amparado por el artículo 18.4 CE.

El RGPD en el considerando 6 se hace eco de la revolución que trae consigo la tecnología en el tratamiento de datos de carácter personal:

“La rápida evolución tecnológica y la globalización han planteado nuevos retos para la protección de los datos personales. La magnitud de la recogida

y del intercambio de datos personales ha aumentado de manera significativa. La tecnología permite que tanto las empresas privadas como las autoridades públicas utilicen datos personales en una escala sin precedentes a la hora de realizar sus actividades. Las personas físicas difunden un volumen cada vez mayor de información personal a escala mundial. La tecnología ha transformado tanto la economía como la vida social, y ha de facilitar aún más la libre circulación de datos personales dentro de la Unión y la transferencia a terceros países y organizaciones internacionales, garantizando al mismo tiempo un elevado nivel de protección de los datos personales".

A lo largo del ciclo del tratamiento de datos por la IA nos podemos encontrar con distintas fases, en las que pueden tratarse datos de carácter personal. Por ello, lo primero que corresponde discernir es si en las distintas etapas de tratamiento se van a tratar o no datos de carácter personal, ya que, en virtud de ello, el tratamiento se sujetará –o no- a la normativa sobre protección de datos personales.

Ejemplo de IA que no utiliza datos de carácter personal: un sistema de IA entrenado con registros de electrocardiogramas previamente anonimizados (se ha eliminado nombre, DNI, fecha de nacimiento, número de historia clínica, dirección y cualquier identificador directo o indirecto con la persona). La IA solo analiza la señal eléctrica del corazón sin saber a qué paciente pertenece. La IA detecta patrones anormales del ritmo cardiaco, clasifica los tipos de arritmia y genera una alerta automática para revisión médica. La salida del sistema de IA podría ser: "probabilidad del 92% de fibrilación auricular detectada en este registro". Todo ello sin ningún dato personal asociado.

Ejemplo de IA que utiliza datos de carácter personal: un sistema de IA que está integrado en la historia clínica electrónica del hospital y que analiza datos personales de un paciente para calcular su riesgo de sufrir un evento cardiovascular. El sistema procesa datos personales (nombre y apellidos, edad, DNI o número de historia clínica, dirección, antecedentes clínicos, resultados de análisis, medicación actual, hábitos, etc.). La IA analiza el perfil completo del paciente y calcula la probabilidad individual de riesgo a 5 o 10 años, recomienda intervenciones personalizadas y genera alertas automáticas. La información de salida podría ser: "Paciente con riesgo cardiovascular alto (23% a 10 años). Se recomienda ajuste de tratamiento antihipertensivo". En este caso, a diferencia del anterior, la información permite identificar

directa o indirectamente a la persona, se vincula el resultado a una historia clínica concreta y se toman decisiones personalizadas individualizadas.

Un sistema de IA basado en aprendizaje automático -*machine learning*- puede tratar datos personales en sus distintas etapas a lo largo del proceso de aprendizaje automático y decisorio.

- Fase de entrenamiento: en la que se le proporcionan a la máquina ingentes cantidades de datos.

- Fase de validación: su objetivo es comprobar si el modelo funciona a nivel experimental.

- Fase de explotación: el sistema realiza inferencias, adopta decisiones e incluso evoluciona a consecuencia de los datos y procesos anteriores.

- Fase de retirada del sistema: bien porque quede obsoleto o bien, simplemente, porque no se utilice.

Así, en la fase de entrenamiento se emplean características que son las propiedades de la información que se trata de aprender. En el caso de los pacientes se incluye información demográfica, como la edad, la historia de la enfermedad, los datos clínicos -exploración física o pruebas complementarias- u otros resultados de interés de la enfermedad, como la respuesta al fármaco o la supervivencia del paciente. En un primer lugar, se utilizan algoritmos de selección con el objeto de identificar qué características contribuyen más a la predicción o clasificación. Este proceso es concretamente el de selección de características. Cuando más dimensiones se añadan más se diferencian los datos y el sistema puede clasificarlos con mayor precisión.

Tras seleccionar las características y el algoritmo hay que disponer de datos de entrenamiento y de evaluación del modelo de variadas fuentes para comprobar si los resultados del algoritmo son reproducibles en distintos entornos.

En la fase de entrenamiento se utilizan las técnicas de aprendizaje automático, bien sea el método supervisado, en cuyo caso se requiere un etiquetado previo de los datos, o no supervisado, en el que la máquina agrupa según características comunes.

Una vez transcurrida la fase de entrenamiento, se prueba el rendimiento de los algoritmos en la muestra de entrenamiento con otro grupo de datos y se elige el que mejor rendimiento proporcione o se sintetizan varios algoritmos para incrementar la eficiencia de la predicción.

No es sencillo determinar si la IA trata datos de carácter personal en cada uno de los procesos del ciclo. Un mayor tratamiento de datos anónimos implica dificultades de validación, de control y dota al sistema de una mayor opacidad. Como contrapunto, se garantiza, en mayor medida, la intimidad personal. Con todo, esta afirmación no resulta categórica.

El tratamiento de datos anónimos implica la no sujeción a la normativa sobre protección de datos, aunque es necesario asegurarse de que la anonimización es plenamente efectiva y evaluar el riesgo de reidentificación existente.

Ejemplo: un sistema de IA analiza datos "anónimos" de *wearables* y movilidad para estudiar la relación entre ritmo cardíaco, calidad de sueño, nivel de actividad física, patrones de movilidad (GPS). Los datos se extraen de pulseras inteligentes y apps de salud. Con carácter previo se eliminan datos como el nombre y apellidos, el correo electrónico y el teléfono. Se sustituye al usuario por un identificador aleatorio y se concluye con que el tratamiento de datos por la IA es anonimizado. La IA detecta patrones de estrés crónico, identifica riesgo cardiovascular, genera perfiles de conductas de salud. Todo ello dentro de un análisis estadístico con datos agregados. Pero ¿se podría reidentificar a la persona? La IA puede detectar el lugar donde la persona duerme cada noche (probable domicilio), el lugar donde pasa ocho horas diarias (probable trabajo), entre otros. Si se cruza con esta información los datos catastrales o padrón, el lugar de trabajo -fácilmente accesible a través de redes sociales como *Linkdln*-, otras redes sociales y datos abiertos, podrá inferirse la identidad con una alta probabilidad. En definitiva, con unos cuantos datos espacio temporales se podría identificar fácilmente a cualquier persona, aunque ni siquiera la identificación esté entre sus finalidades. Conforme a la normativa, si existe una posibilidad razonable del reidentificación los datos siguen siendo personales y esto en el entorno de trabajo de una IA avanzada es muy habitual, de modo que en la era de la IA muchos conjuntos "anonimizados" de datos, dejan de ser anónimos al analizar patrones complejos y al cruzarse con distintas bases de datos, por lo que les será de aplicación todo cuanto vamos a tratar aquí con relación a los datos de carácter personal utilizados por sistemas de IA.

2. COEXISTENCIA ENTRE LEY DE IA Y RGPD

La Ley de IA es el primer marco jurídico global en materia de IA, que aborda sus riesgos con el objetivo de fomentar una IA fiable en Europa.

La Ley de IA prevé normas desde un enfoque del riesgo para los desarrolladores e implementadores de sistemas de IA con relación a los usos específicos salvaguardando la seguridad, los derechos fundamentales desde una IA antropocéntrica, que, a su vez, refuerce la inversión y la innovación de IA en Europa.

Entre los riesgos germen de esta norma figura el que plantea la puesta en circulación de sistemas de IA en el ámbito específico de la protección de datos de carácter personal, motivo por el cual, ambas normas -el RGPD y la Ley de IA- están en estrecha relación cuando se trata de sistemas de IA que tratan datos de carácter personal.

Del artículo 2.1 del RGPD se extrae la aplicación de la norma al tratamiento de los datos personales, total o parcialmente automatizados y también al tratamiento de datos personales no automatizados almacenados o que vayan a serlo en un fichero. Por lo tanto, a la IA se le aplicará el RGPD cuando se traten datos personales, que según el artículo 4.1 del RGPD y 3.50 de la Ley de IA son toda información que haga referencia a una persona física identificada o identificable.

En este sentido el caso SyRi en los Países Bajos constituyó un precedente, un ejemplo clave de los límites que, desde la normativa, deben establecerse en el uso de la IA.

Syri era una herramienta utilizada por el gobierno de Países Bajos para detectar posibles fraudes en ayudas sociales, impuestos y otros servicios públicos. El sistema cruzaba grandes cantidades de datos personales de distintas administraciones y generaba perfiles de personas consideradas de “alto riesgo”.

El problema era la falta de transparencia y control. Nadie sabía con claridad qué datos se usaban ni cómo se decidía que una persona era sospechosa. Además, SyRI se aplicaba sobre todo en barrios vulnerables, lo que aumentaba el riesgo de discriminación y estigmatización.

En 2020, el Tribunal de Distrito de La Haya declaró ilegal el sistema. El Tribunal consideró que SyRI violaba el derecho a la privacidad protegido por el Convenio Europeo de Derechos Humanos.

El Tribunal no prohibió la IA en general, pero dejó claro que no puede usarse sin límites. Cuando una herramienta tecnológica afecta a derechos fundamentales, debe ser clara, justa y controlable.

Mientras la Ley de IA regula la seguridad de los productos y las tecnologías que tratan datos personales, el RGPD protege los datos personales y garantiza los derechos de la persona cuando sus datos son tratados, especialmente en un entorno digital. Su enfoque es distinto, pero, claramente, se complementan.

Ejemplo:

En un hospital se utiliza un sistema de IA para la detección del cáncer de pulmón a partir de radiografías y pruebas de TAC. La IA analiza miles de imágenes médicas de pacientes y las compara con casos anteriores para ayudar al médico a identificar patrones sospechosos. La decisión final la tomará el profesional, pero la IA es una tecnología que constituye un apoyo a la decisión médica.

Para ello, se tratan datos personales y, concretamente, datos de salud como imágenes médicas, edad, sexo del paciente e historial clínico relevante.

El RGPD es aplicable puesto que se tratan datos personales, y, además, muy sensibles como los son los de salud. Su ámbito estriba en la consideración de las siguientes cuestiones:

- ¿Existe una base legal para el tratamiento de los datos? En este caso está justificado, bien por el interés del tratamiento de datos en beneficio de la salud del paciente, del interés público que reviste el tratamiento en el ámbito sanitario, o por el consentimiento del paciente.

- ¿Qué datos personales se pueden tratar? El sistema de IA únicamente podrá utilizar los datos personales estrictamente necesarios, no más de los imprescindibles para la finalidad determinada.

- ¿Hay que informar al paciente? En virtud del principio de transparencia, el paciente deberá ser informado de aspectos como que sus datos van a ser utilizados por un sistema de IA, la finalidad y quién es el responsable del tratamiento.

- ¿Qué derechos tiene el paciente sobre sus datos personales? El paciente podrá ejercer derechos como el acceso, la rectificación y, en algunos casos y con condiciones, el de supresión.

- ¿Qué obligaciones tiene el hospital? El hospital, entre otras, deberá establecer las medidas técnicas y organizativas oportunas para la protección frente a accesos no autorizados.

En resumen, el RGPD protege al paciente como titular de los datos.

La Ley de IA tiene por objeto no tanto los datos personales, sino el riesgo generado por el sistema de IA.

La IA médica es considerada de riesgo alto porque puede afectar a la salud y a la vida de las personas. Ello implica exigencias reguladas en la Ley de IA como:

- La evaluación y el control del riesgo antes del uso del sistema

- Datos de entrenamiento de calidad para evitar errores y sesgos

- Documentación y trazabilidad para saber cómo funciona el sistema.

- Supervisión humana significativa: la IA no es sustituta del médico.

- Transparencia técnica: imprescindible para que el sistema pueda auditarse.

8 Protección de datos personales y Sistemas de IA

Dra. Cristina Gil Membrado

La Ley de IA controla cómo se diseña y utiliza la IA para que sea segura y fiable.

Ambas normas se complementan para que la IA sea segura, útil y respetuosa con los derechos de la persona.

- RGPD: protege datos personales / Ley de IA: regula el sistema de IA
- RGPD: se centra en la privacidad / Ley de IA: se centra en seguridad y riesgos
- RGPD: protege derechos afectados por el tratamiento de datos personales / Ley de IA: protege derechos afectados por el nivel de riesgo del sistema.

En conclusión:

- El RGPD: regula el tratamiento de datos
- La Ley de IA: regula el funcionamiento y los riesgos del sistema de IA.

Ambos cuerpos normativos se aplican simultáneamente, pero con fines distintos, aunque complementarios.

3. LOS DATOS: EL COMBUSTIBLE QUE HACE POSIBLE LA IA

Para entrenar, desplegar y usar un sistema de IA es necesario tratar datos. Los datos no son solo imprescindibles para que la IA aprenda, sino también para que sea segura en su aplicación y funcionamiento.

3.1. La IA necesita datos de entrada para su entrenamiento

La IA no razona por sí sola. Compara al paciente actual con patrones aprendidos de otros pacientes.

Para ello, previamente hay que nutrir a la IA con datos relativos a historias clínicas previas, imágenes radiológicas etiquetadas, analíticas con diagnóstico final conocido, evoluciones clínicas y desenlaces entre muchos otros.

Los datos de entrenamiento son los "pacientes anteriores" con los que la IA ha aprendido medicina. La Ley de IA los define en el artículo 3.29 como los datos usados para entrenar un sistema de IA mediante el ajuste de sus parámetros entrenables.

En el caso de un sistema de aprendizaje supervisado, se tratará normalmente de un conjunto de ejemplos que emparejan algunos datos de entrada con el resultado esperado.

Ejemplo: sistema de IA que detecta neumonías en radiografías. Se le proporcionan miles de radiografías clasificadas por el médico como "normal" o "neumonía". La IA aprende patrones, lo que implica detectar qué rasgos en la imagen aparecen cuando hay infección. A partir de aquí realiza predicciones nuevas al recibir una nueva radiografía estimando la probabilidad de neumonía.

En los sistemas de aprendizaje no supervisado, no se proporcionan resultados esperados (diagnósticos previos), sino los datos de entrada simplemente (la IA ve muchas radiografías y busca patrones por sí misma).

Ejemplo: sistema de IA que observa miles de imágenes sin saber cuáles son normales o patológicas. Agrupa por similitud ya que detecta que ciertas

imágenes se parecen entre sí y forma grupos (a modo de ejemplo, un grupo con pulmones claros y otro con opacidades). También detecta anomalías, es decir, si aparece una radiografía muy diferente al resto la marca como inusual para que el médico la revise de modo prioritario. Esta IA no dice "es neumonía" sino "esta imagen es diferente a lo habitual". Este sistema de IA es útil como sistema de alerta temprana o para descubrir patrones nuevos, apoyando al profesional sanitario sin necesidad de datos previamente etiquetados.

Por último, en los sistemas de aprendizaje reforzado, no se facilitan respuestas esperadas, pero se debe suministrar al sistema información sobre la rentabilidad de las diferentes opciones.

Ejemplo: aquí el sistema no solo "mira" y "decide" sino que aprende una estrategia como si fuera un residente en formación. La IA observa la radiografía, decide qué zona mirar primero (por ejemplo, el lóbulo inferior), puede ampliar a otra región si tiene dudas y finalmente emite un diagnóstico (si hay o no neumonía). Después recibe una recompensa ya que gana puntos si acierta, pierde muchos puntos si se equivoca (no detecta una neumonía) y pierde puntos -aunque menos- si tarda demasiado o revisa zonas innecesarias. La IA aprende mediante este método y mediante el análisis de miles de radiografías qué estrategia da más puntos, qué áreas mirar antes como típicas, a no precipitarse y a evitar errores graves. Por lo tanto, no solo aprende a reconocer patrones en la imagen sino también a tomar decisiones paso a paso, optimizando la precisión y la seguridad clínica.

La IA no conoce fisiopatología, sino que aprende correlaciones clínicas a partir de los datos de entrenamiento.

Los datos de entrenamiento sirven:

- Para aprender lo que es normal y lo que es patológico: la IA aprende qué aspecto suele tener una TAC "normal", qué constantes preceden a una sepsis, y qué patrones se asocian a una mala evolución. No aprende literatura médica, sino frecuencias y asociaciones.

- Para aprender a priorizar riesgos: por ejemplo, tras proporcionarle datos de miles de pacientes con dolor torácico se entrena con quién tuvo IAM y quién no y aprende qué combinaciones aumentan o reducen el riesgo.

- Para imitar decisiones humanas previas: a partir de diagnósticos hechos por médicos, de decisiones terapéuticas o de criterios de alta o de ingreso previos. La IA reproducirá en estos casos la práctica médica previa, con sus aciertos y con sus errores.

3.2. La IA necesita datos para ser validada antes de usarse en pacientes

La IA necesita datos para comprobar si funciona para lo que ha sido diseñada y formada, para generalizar con pacientes distintos de los que constituyen los datos de entrada y si comete errores clínicamente relevantes.

La IA debe ser validada antes de usarse en pacientes porque sus decisiones influyen en la salud y sin validación hay un gran riesgo de que cometa errores, genere diagnósticos erróneos o discrimine a grupos de pacientes.

La validación implica probarla con datos reales y representativos antes de aplicarla en la práctica clínica para comprobar: si sus predicciones son precisas y fiables, si funciona bien con todo tipo de pacientes (edad, sexo, patologías, contexto), si no introduce sesgos que pueden perjudicar o discriminar a grupos y si sus resultados son comparables, como mínimo, a los de los profesionales de la salud.

Primero, la IA se entrena con miles de historiales clínicos. Después, se prueba con un conjunto de validación, que son datos de otros pacientes reales que el modelo no ha visto antes. Si en este grupo nuevo la IA sigue prediciendo correctamente que el sistema funciona bien más allá del entrenamiento.

La Ley de IA define en el artículo 3.30 datos de validación como los datos usados para proporcionar una evaluación del sistema de IA entrenado y adaptar sus parámetros no entrenables y su proceso de aprendizaje para, entre otras cosas, evitar el subajuste o el sobreajuste. En el artículo 3.31 define conjunto de datos de validación como un conjunto de datos independiente o una parte del conjunto de datos de entrenamiento, obtenida mediante una división fija o variable.

Los datos de validación se utilizan para ajustar el modelo entrenado, lo que permite a los creadores del modelo elegir entre diferentes procesos y estrategias de aprendizaje. Por ejemplo, propicia a los creadores evitar el

fenómeno de sobreajuste, en el que un modelo aprende reglas que describen bien el conjunto de entrenamiento, pero que no se generalizan correctamente.

Ejemplo: el sistema de IA para detectar neumonía en radiografías aprende durante el entrenamiento con radiografías que provenían de un hospital que usaba un marcador digital en la esquina de la imagen. La IA aprendió que "si aparece ese pequeño texto en la esquina: hay neumonía". En las pruebas de validación en el hospital el sistema arrojaba unos resultados excelentes (95% de aciertos), pero al ser usado en otro hospital donde no existía ese marcador, el rendimiento cae en picado. Nos encontramos con un sobreajuste -*overfiftting*- en que la IA no aprendió realmente de los patrones médicos de la neumonía (como opacidades pulmonares) sino de detalles accidentales suministrados por los datos de entrenamiento. El modelo memorizó peculiaridades de los datos de entrenamiento en vez de aprender señales clínicas reales, por lo que falla en el mundo real.

Habitualmente se utilizan datos históricos anonimizados, al margen de los datos de entrenamiento para evaluar su desempeño.

Ejemplo:

- Una IA diseñada para detectar el cáncer de pulmón en radiografías.
- El sistema se entrena con miles de radiografías diagnosticadas por médicos.
- Antes de usarla con pacientes reales, se valida con nuevas radiografías cuyos diagnósticos se conocen a los que la IA es ajena.
- Se compara lo que predice la IA con el diagnóstico médico real.
- Si la IA alcanza una alta precisión y una baja tasa de errores puede aprobarse para pruebas clínicas controladas
- Después de superar estas fases se utilizará como apoyo a la decisión médica, nunca como sustituto.

Sin datos para validar, no podríamos tener garantía de que la IA es segura y efectiva para el paciente.

3.3. Datos de prueba

Para evaluar el rendimiento final de un sistema de IA se utilizan datos de prueba, que son conjuntos de datos clínicos usados una vez el sistema ya ha sido entrenado y validado con el objeto de comprobar si funciona correctamente en situaciones reales antes (o durante) su uso en la práctica clínica.

- No se utilizan para entrenar el modelo
- Son independientes de los datos de entrenamiento y de validación
- Representan casos clínicos reales y variados
- Miden la precisión, los errores, los sesgos y la seguridad clínica
- Son fundamentales para determinar si la IA es fiable y segura para los pacientes

Ejemplo:

- Una IA está diseñada para predecir la sepsis en pacientes hospitalizados.
- La IA se ha entrenado con historiales clínicos de pacientes pasados
- La IA se ha validado con otro conjunto distinto de datos para ajustar su funcionamiento
- La IA se prueba con datos como historias clínicas reales de otros hospitales y de distintos periodos.
- Se compara la predicción de la IA con el hecho de que el paciente desarrollara sepsis o no.

- Si la IA mantiene una alta precisión podrá considerarse como herramienta de apoyo clínico.

Los datos de prueba permiten verificar si una IA médica generaliza correctamente y es segura antes de su uso en pacientes reales.

Suelen ser datos de salud anonimizados o seudonimizados.

Un dato anonimizado es aquel que ha sido tratado de modo que ya no es posible identificar a la persona a la que se refiere ni directa ni indirectamente. Si bien el RGPD no define formalmente la anonimización, sí aclara en el Considerando 36 que "Los principios de protección de datos no se aplican a la información anónima, es decir, aquella que no guarda relación con una persona física identificada o identificable o que se ha anonimizado de forma irreversible". Ello implica que la identificación de la persona es irreversible, que no existe ninguna información adicional que permita reidentificar al interesado y que no es posible identificar a alguien ni combinando el dato con otra información razonablemente disponible. Una vez anonimizado correctamente, deja de ser dato personal y el RGPD ya no aplica.

Ejemplo: antes María López, 34 años, Palma de Mallorca, historial médico. Después (anonimizado): Paciente mujer, grupo de edad 30-40, diagnóstico X dentro de un conjunto estadístico. Si no hay forma razonable de volver a saber quién es la persona, el dato está anonimizado.

Un dato seudonimizado es aquel que ha sido tratado de forma que ya no puede atribuirse a una persona concreta sin utilizar información adicional. La definición conforme al artículo 4.5 del RGPD de seudonimización es "el tratamiento de datos personales de manera tal que ya no puedan atribuirse a un interesado sin utilizar información adicional, siempre que dicha información adicional figure por separado y esté sujeta a medidas técnicas y organizativas destinadas a garantizar que los datos personales no se atribuyan a una persona física identificada o identificable."

Los datos siguen siendo datos personales (no dejan de estar protegidos por el RGPD). Se sustituye la información identificativa (por ejemplo, nombre o DNI) por un código o identificador que es la "clave" que permite volver a identificar a la persona u se guarda por separado protegida con medidas de seguridad adecuadas.

Ejemplo: antes María López — DNI 12345678A — historial médico. Después (seudonimizado): Paciente XJ-204 — historial médico. Si existe una tabla aparte que relaciona "XJ-204" con María López, el dato está seudonimizado, no anonimizado.

3.4. Datos de entrada

La Ley de IA también introduce la categoría de datos de entrada, que, según el artículo 3.33, se trata de datos que se proporcionan a un sistema de IA o que este adquiere directamente y sobre cuya base el sistema produce una salida. Estos datos pueden estar contenidos en la solicitud del usuario, el *prompt*, o proceder de una cuenta de usuario a la que el sistema de IA tiene acceso.

Ejemplo de datos de entrada vía *prompt* en una IA médica conversacional: paciente mujer, 67 años. Fiebre 38,5 °C desde hace 3 días, tos productiva, dolor torácico al respirar, saturación de oxígeno 91%, antecedente de EPOC. ¿Cuál es el diagnóstico más probable y qué pruebas recomendar?". En este caso los datos de entrada no son una imagen ni una señal biomédica directa sino información clínica escrita en lenguaje natural que consta de edad, síntomas, constantes vitales y antecedentes. La IA a partir de esta petición podrá sugerir diagnósticos probables (por ejemplo, neumonía), o recomendar pruebas (radiografía de tórax, analítica, entre otras) o señalar factores de riesgo. En definitiva, el *prompt* actúa como la ficha clínica digital que alimenta al sistema para generar una orientación médica.

Los datos de entrada no siempre consisten en un texto introducido manualmente.

Ejemplo: sistema de IA que detecta arritmias cardíacas. En este supuesto los datos de entrada son: señales eléctricas del corazón (ECG), frecuencia cardíaca en tiempo real, intervalos entre latidos y edad y sexo del paciente. La IA recibe directamente la señal digital del ECG, analiza patrones (ritmo irregular, ausencia de ondas P, variaciones en los intervalos) y determina si existe, a modo de ejemplo, una fibrilación auricular. En este caso la entrada es a partir de dato biomédico procesado automáticamente por el sistema (señales, imágenes o datos estructurados).

Los datos de entrada es la información clínica que se introduce en el sistema de IA cuando ya se encuentra en funcionamiento para que genere una

predicción, recomendación o apoyo a la decisión médica sobre un paciente concreto.

Los datos de entrada:

- Se usan en tiempo real o en práctica clínica.
- Corresponden a pacientes actuales, no históricos.
- Determinan directamente la salida clínica de la IA.
- Son casi siempre datos personales y datos de salud.
- Su calidad afecta de forma inmediata a la seguridad del paciente.

En IA médica, los datos de entrada deben ser exactos, completos y actualizados.

Ejemplo:

En una IA de apoyo al diagnóstico de insuficiencia cardiaca serán datos de entrada:

- Edad y sexo del paciente
- Presión arterial
- Frecuencia cardiaca
- Resultados de analíticas
- Ecocardiograma
- Datos registrados en la historia clínica

La IA procesará los datos y proporcionará una información de salida que podrá consistir en la probabilidad de insuficiencia cardiaca o en una alerta de riesgo para el médico.

Si los datos de entrada son incompletos o erróneos (por ejemplo, un valor mal registrado), la salida de la IA puede ser incorrecta y afectar al tratamiento del paciente.

Los datos de entrada son los datos clínicos del paciente que alimentan a la IA durante su utilización habitual para apoyar al médico en su decisión y que a la vez alimentarán y enriquecerán el aprendizaje del sistema de IA para nuevos casos.

Ejemplo:

Sistema de IA que detecta retinopatía diabética a partir de imágenes de fondo de ojo. La IA analiza la imagen y emite un resultado: "probable retinopatía moderada". El oftalmólogo revisa el caso y confirma el diagnóstico real. Esa confirmación (correcto o incorrecto) se guarda en el sistema y periódicamente el modelo se vuelve a entrenar incorporando esos nuevos casos ya validados.

Si la IA se equivoca, el sistema aprende de ese error. Si la IA acierta, refuerza ese patrón como válido.

A través de ello la IA mejora progresivamente la precisión, se adapta a nuevos dispositivos de cámara o poblaciones y reduce errores con el tiempo. Se trata de una IA no estática, sino que va a aprendiendo de la práctica clínica real utilizando la retroalimentación para una mejora continua.


AI

4. EL TRATAMIENTO DE DATOS EN LA LEY DE IA Y EN EL RGPD

Un hospital implanta un sistema de IA que analiza mamografías para detectar indicios tempranos de cáncer de mama y ayuda al radiólogo a tomar decisiones.

El sistema:

- Analiza imágenes médicas del paciente.
- Está integrado en la historia clínica electrónica.
- Genera una probabilidad diagnóstica.
- Sugiere prioridad de atención.

¿Cómo actúa el RGPD?

El sistema trata:

- Datos identificativos personales.

- Datos de salud personales que forman parte de una categoría especialmente sensible.

Las obligaciones principales del hospital serán:

- Tratar los datos bajo una base jurídica válida (prestación de asistencia sanitaria o cumplimiento de una obligación legal).
- Tratar los datos con las restricciones establecidas para las categorías espaciales de datos.
- Realizar una evaluación de impacto (EIPD).

- Principio de minimización: utilizar solo los datos estrictamente necesarios.
- Implantar medidas técnicas y organizativas (cifrado, control de accesos y registro de actividades).
- Contemplar los derechos de los pacientes (acceso, rectificación, limitación, oposición, supresión -cuando proceda- e información sobre decisiones automatizadas).

¿Cómo actúa la Ley de IA?

El sistema es de alto riesgo porque se utiliza en el ámbito sanitario, influye en las decisiones diagnósticas y puede afectar a la salud y a la seguridad de las personas.

- Las obligaciones para el hospital son:

    - Para el proveedor del sistema:
    - Sistema de gestión de riesgos.
    - Gobernanza de datos de entrenamiento.
    - Documentación técnica detallada.
    - Registro en base de datos europea.
    - Evaluación de conformidad.
    - Marcado CE.

- Para el hospital:

    - Supervisión humana obligatoria.
    - Uso conforme a instrucciones.
    - Monitoreo de funcionamiento.
    - Notificación de incidentes graves.

La Ley de IA y el RGPD desempeñan un papel esencial en la regulación del tratamiento de datos por los sistemas de IA a lo largo de todo su ciclo de vida, desde el desarrollo hasta su uso real. Ambos marcos normativos son especialmente relevantes cuando la IA se aplica en ámbitos sensibles como la salud, donde los errores pueden tener consecuencias graves para las personas. Nos detendremos exclusivamente en los preceptos principales relacionados con el tratamiento de datos personales por los sistemas de IA.

En primer lugar, hemos visto que los datos de entrenamiento son aquellos utilizados para enseñar al sistema de IA a reconocer patrones. La importancia de la Ley de IA radica en que exige que estos datos sean de calidad, representativos y adecuados al fin previsto, con el objetivo de evitar sesgos y resultados discriminatorios. Por su parte, el RGPD es clave porque muchos datos de entrenamiento son datos personales, incluidos datos de salud, lo que obliga a contar con una base jurídica válida, respetar los principios de minimización y exactitud, y aplicar medidas de seguridad. Sin estas garantías, el entrenamiento de la IA podría vulnerar derechos fundamentales.

En segundo lugar, los datos de validación se utilizan para ajustar el funcionamiento del sistema y comprobar si aprende correctamente. Aquí, la Ley de IA refuerza la necesidad de un control técnico riguroso, asegurando que el sistema no solo funciona, sino que lo hace de manera fiable y conforme a los riesgos identificados. El RGPD vuelve a ser relevante porque el tratamiento de datos personales en esta fase debe seguir siendo lícito y proporcional, evitando usos secundarios no autorizados.

En tercer lugar, los datos de prueba sirven para evaluar el rendimiento final del sistema antes de su despliegue o durante auditorías posteriores. La Ley de IA es fundamental porque vincula estos datos a la gestión de riesgos y a la

demostración de conformidad, garantizando que el sistema es seguro antes de su uso con pacientes reales. Desde la perspectiva del RGPD, los datos de prueba deben tratarse con las mismas garantías que cualquier otro dato personal: finalidad específica, seguridad y, cuando proceda, evaluaciones de impacto en protección de datos.

Por último, los datos de entrada son los datos reales de los pacientes introducidos en la IA durante su uso clínico. En esta fase, la importancia del RGPD es máxima, ya que se trata de datos de salud actuales que afectan directamente a la persona. El Reglamento garantiza derechos como la protección frente a decisiones exclusivamente automatizadas y la exigencia de exactitud y confidencialidad. La Ley de IA complementa esta protección al exigir supervisión humana, robustez y control del riesgo, evitando que la IA sustituya al profesional sanitario.

En definitiva, la Ley de IA y el RGPD actúan de forma complementaria. La primera asegura que los sistemas de IA sean técnicamente seguros y fiables, mientras que el segundo protege los derechos fundamentales de las personas cuyos datos se utilizan. Ambos son indispensables para generar confianza, seguridad jurídica y protección efectiva en el uso de la IA.

Veamos ejemplos de la incidencia de ambas normas.

4.1. Ley de IA

- **Artículo 9:**

  Gestión de riesgos. Regula la identificación y mitigación de riesgos en todo el ciclo de vida del sistema. Obliga a analizar los riesgos derivados de datos de entrenamiento sesgados o defectuosos y a evaluar los riesgos clínicos antes de usar la IA con pacientes reales. Los resultados obtenidos con los datos de prueba sirven para detectar riesgos residuales (errores diagnósticos, falsos negativos, etc.). Los riesgos derivados de datos de entrada incorrectos (p. ej., valores clínicos mal registrados) deben ser identificados y mitigados.

El Artículo 9 exige que la IA médica no solo funcione, sino que identifique, evalúe y reduzca riesgos desde el diseño hasta su uso real con pacientes, incluyendo los derivados de datos sesgados, errores diagnósticos o entradas incorrectas.

Ejemplo:

Imaginemos una IA diseñada para detectar cáncer de piel a partir de fotografías de lesiones cutáneas. Se descubre que la mayoría de las imágenes utilizadas para entrenar la IA eran de personas de piel clara. El riesgo en este caso es que falle en personas con piel oscura. La medida será reentrenar el modelo incorporando imágenes diversas. El riesgo también puede provenir de la detección de demasiados falsos negativos (cáncer no detectado) en cuyo caso habrá que ajustar el umbral de sensibilidad y obligar a revisión médica en los casos mínimamente dudosos. El riesgo también puede provenir de datos de entrada consistentes en fotos borrosas por lo que la IA pueda equivocarse. En este caso habrá que establecer un mecanismo de bloqueo del análisis si la calidad de la imagen no es suficiente.

- **Artículo 10:**

  Datos y gobernanza de datos. Regula los requisitos de calidad, representatividad y gobernanza de los datos. Obliga a que los datos de entrenamiento sean relevantes, representativos, libres de errores y adecuados para el fin médico previsto. Impone la validación con datos fiables antes del uso clínico. Los datos de prueba deben permitir comprobar que la IA funciona correctamente antes de su uso clínico, evitando sesgos o fallos peligrosos para los pacientes. El sistema debe estar diseñado para gestionar adecuadamente datos de entrada incompletos, erróneos o atípicos, minimizando riesgos clínicos.

El Artículo 10 exige que la IA médica se base en datos fiables, representativos y bien gestionados, y que esté preparada para manejar errores o datos incompletos sin poner en riesgo a los pacientes.

Ejemplo:

Imaginemos una IA que solo ha sido entrenada con datos de adultos jóvenes, podría fallar en personas mayores. La medida a adoptar pasa por usar datos variados (edad, sexo, distintos centros sanitarios) y revisar que no contengan errores en valores clínicos. También habrá que probarla con datos fiables y separados de los del entrenamiento. Habrá que realizar pruebas para detectar sesgos o fallos de modo que una vez se detecta con qué población falla (enfermos crónicos) habrá que corregir el modelo. También resulta fundamental la gestión de datos incompletos o erróneos. Si falta un valor

importante (lactato) o hay un dato claramente imposible (temperatura de 55º C) el sistema debe detectarlo y no generar una predicción automática sin advertencia.

- **Artículo 11:**

  Documentación técnica. Regula la documentación sobre el diseño, el desarrollo y la validación del sistema. Obliga a incluir el origen, la naturaleza y el tratamiento de los datos de entrenamiento y permite demostrar que la IA ha sido correctamente validada. Permite demostrar ante autoridades que la IA fue correctamente evaluada antes de su uso.

El Artículo 11 no trata de cómo funciona la IA en sí, sino de que exista prueba documentada y verificable de que fue diseñada, entrenada y validada de forma rigurosa antes de su uso clínico.

Ejemplo:

Para que una IA que analiza radiografías de tórax para detectar neumonía se implante en hospitales el fabricante debe tener un expediente técnico que acredite que su diseño y validación estén claramente documentados. En el expediente deberá constar: cómo se diseñó el sistema (tipo de modelo utilizado; objetivo clínico concreto); origen y tratamiento de los datos (de qué hospitales proceden las radiografías; cuántas imágenes se usaron; cómo se anonimizaron; cómo se etiquetaron -quién confirmó el diagnóstico-); cómo se validó (qué resultados obtuvo en datos de prueba independientes; tasa de errores, falsos negativos y falsos positivos; en qué condiciones puede fallar). Por medio de esta documentación, el hospital y las autoridades sanitarias pueden comprobar que la IA fue evaluada correctamente antes de ser usada con pacientes, que sus datos son adecuados y que sus límites y sus riesgos están identificados.

- **Artículo 13:**

  Transparencia. Regula la información clara para los usuarios del sistema. Obliga a que el profesional sanitario conozca las limitaciones derivadas de los datos de entrenamiento. Información clara para usuarios y profesionales sanitarios. El médico debe conocer los límites y fiabilidad de

la IA validada. El médico debe ser informado de qué tipo de datos de entrada necesita la IA y cómo afectan a los resultados.

El Artículo 13 garantiza que la IA médica no sea una "caja negra" para el profesional, sino una herramienta cuyos límites, fiabilidad y requisitos de uso estén claramente explicados.

Ejemplo:

Una IA ayuda a detectar fracturas en radiografías en un servicio de urgencias. Cuando el médico usa la IA, el sistema debe indicar claramente: qué hace y qué no hace (si está validada para fracturas en adultos o también en niños o si diagnostica); sus limitaciones (si fue entrenada con radiografías de buena calidad o si puede fallar más en imágenes borrosas o en fracturas pequeñas); su nivel de fiabilidad (si tiene una sensibilidad del 94% o si tiene riesgo de falsos negativos en casos determinados); qué datos necesita y cómo influyen (requiere radiografías en determinadas proyecciones o si la imagen está incompleta o mal posicionada, en cuyo caso la precisión disminuye). Todo ello es para que el médico decida cuándo puede confiar más o menos en la herramienta, cuándo extremar la revisión o en qué situaciones puede no ser adecuada.

- **Artículo 14:**

  Supervisión humana. Regula el control humano significativo sobre la IA. Implica la reducción de riesgos cuando el entrenamiento no ha cubierto de modo efectivo los supuestos clínicos. Impide que la IA tome decisiones clínicas sin supervisión humana. El profesional sanitario debe poder revisar, corregir o ignorar la salida de la IA cuando los datos de entrada no sean fiables.

El Artículo 14 garantiza que la IA médica funcione como herramienta de apoyo bajo supervisión humana, evitando decisiones automáticas sin control y permitiendo que el profesional revise, corrija o descarte sus resultados cuando sea necesario.

Ejemplo:

Una IA calcula el riesgo de ictus en pacientes hospitalizados a partir de constantes vitales y de antecedentes clínicos. La IA no decide de modo automático, sino que muestra "riesgo alto de ictus en 24 horas" sin activar tratamiento ya que la decisión la toma el médico, quien puede revisar y corregir. El médico comprueba los datos y detecta que la presión arterial estaba mal registrada. En este caso corregirá el dato y volverá a generar la evaluación o bien ignorará la recomendación. También puede suceder que el paciente tenga una condición rara que apenas aparecía en los datos de entrenamiento. En este caso el sistema debe advertir que la fiabilidad puede ser menor. La IA actúa como apoyo, no como sustituta y es el profesional el que mantiene el control clínico en todo momento.

- **Artículo 15:**

  Exactitud, solidez y ciberseguridad. Exige que la IA sea robusta frente a errores. El sistema debe mantener un rendimiento seguro incluso cuando los datos de entrada sean imperfectos.

El Artículo 15 garantiza que la IA médica no solo sea precisa en condiciones ideales, sino también resistente a errores técnicos, datos imperfectos y amenazas de seguridad, manteniendo un rendimiento seguro para los pacientes.

Ejemplo:

Una IA detecta hemorragias cerebrales en TAC en el servicio de urgencias. Antes de usarse, habrá que comprobar su exactitud, de modo que la IA demuestre un alto nivel de sensibilidad y especificidad en estudios de validación. Por otro lado, si el TAC tiene algo de ruido, pequeñas variaciones técnicas o un contraste diferente al habitual, el sistema debe seguir funcionando con seguridad o, si la calidad no es suficiente, advertirlo en lugar de dar un resultado poco fiable. Si la IA recibe un archivo incompleto o dañado no debe generar un diagnóstico erróneo sino bloquear el análisis y notificar el problema. Con relación a la ciberseguridad debe estar protegido frente a accesos no autorizados o manipulaciones de datos que pudieran alterar sus resultados.

- **Artículo 47:**

  Declaración de conformidad. Regula la verificación previa a la comercialización. Obliga a la inclusión de evidencias de que los datos de entrenamiento cumplen los requisitos normativos. Incluye la evidencia de la validación previa del sistema. La conformidad depende, entre otros aspectos, de que los datos de prueba evidencien un rendimiento seguro y fiable.

El Artículo 47 garantiza que una IA médica no pueda ponerse en el mercado sin antes demostrar, con pruebas documentadas, que es segura, validada y conforme a la normativa aplicable.

Ejemplo: una empresa ha desarrollado una IA para detectar embolia pulmonar en TAC y quiere comercializarla en hospitales. Antes de salir al mercado, el fabricante deberá declarar formalmente que el sistema cumple todos los requisitos legales. Para ello tendrá que demostrar que los datos de entrenamiento cumplen la normativa (son adecuados para el objetivo clínico, son representativos y han sido correctamente gestionados y documentados); existe validación previa suficiente (el sistema ha sido probado con datos independientes, se ha medido la sensibilidad, la especificidad y la tasa de errores); los datos de prueba evidencian seguridad y fiabilidad (el rendimiento es estable y no presenta fallos graves que pongan en riesgo a los pacientes). Con estas evidencias el fabricante emite la declaración de conformidad afirmando que la IA ha sido evaluada y cumple con los requisitos exigidos para su comercialización.

4.2. RGPD

- **Artículo 5:**

  Principios del tratamiento. Regula la licitud, la minimización, la exactitud y la limitación de la finalidad. Obliga a que los datos de entrenamiento sean adecuados, exactos y limitados a lo necesario. Obliga a que los datos usados para validar la IA deben ser correctos y pertinentes. Los datos de prueba deben ser adecuados y necesarios para evaluar la IA, sin usar más datos personales de los imprescindibles. Los datos de

entrada deben ser correctos, actuales y estrictamente necesarios para la finalidad clínica de la IA.

El Artículo 5 garantiza que la IA médica trate los datos personales de forma lícita, exacta y limitada a lo necesario para su finalidad clínica, evitando usos excesivos o inadecuados de información sensible.

Ejemplo:

Una IA predice el riesgo de complicaciones tras una cirugía. Será necesario que los datos sean adecuados y limitados en el entrenamiento (para predecir complicaciones quirúrgicas la IA necesita tratar datos de edad, antecedentes médicos y resultados analíticos relevantes). No sería lícito incluir datos como la información financiera del paciente, y, al contrario, solo se recabarán los datos estrictamente necesarios para la finalidad clínica. También será necesario que los datos de validación sean exactos. Para ello, antes de validar el sistema se revisará que los historiales clínicos estén correctamente registrados y si hubiera errores diagnósticos o de otro tipo se corregirán para no sesgar los resultados. Los datos de prueba deberán ser proporcionales. Ello implica que, para evaluar el rendimiento, se utilizarán únicamente los datos imprescindibles para comprobar su seguridad y eficacia, evitando recopilar más información personal de la necesaria. En lo referente a los datos de entrada, deberán ser correctos y actuales y si estuvieran incompletos o no actualizados el sistema deberá advertirlo.

- **Artículo 6:**

  Licitud del tratamiento. Regula las bases jurídicas del tratamiento de datos personales. Obliga a que el uso de los datos de los pacientes para entrenar la IA tenga una base legal válida. Obliga a la obtención de base legal -por ejemplo, el consentimiento- para la validación con datos personales de pacientes. El uso de datos de pacientes como datos de prueba solo es lícito si existe consentimiento, base legal sanitaria o investigación científica. El uso de datos de entrada para IA clínica debe apoyarse en asistencia sanitaria, obligación legal o consentimiento, según el caso.

El Artículo 6 garantiza que los datos de salud no puedan utilizarse para entrenar, validar o usar una IA médica sin una base legal clara y legítima,

como el consentimiento o una habilitación sanitaria o científica prevista en la ley.

Ejemplo:

Una IA predice el riesgo de recaída en pacientes oncológicos usando historiales clínicos. En primer lugar, y en lo relativo al entrenamiento del sistema, el hospital requiere usar miles de historiales médicos para entrenar a la IA (solo podrá hacerlo si existe una base legal válida como el consentimiento del paciente, su uso para investigación científica conforme a la ley o una base jurídica prevista en la normativa). Deberá validarse con datos reales por lo que antes de implantarla deberá ser alimentada con datos de otros pacientes (deberá existir una base legal como la investigación biomédica autorizada o el consentimiento). Para usar datos de prueba deberá justificarse jurídicamente que ese uso es necesario y lícito. Con respecto a los datos de entrada en la práctica clínica cuando el médico introduce datos actuales del paciente para obtener una predicción el tratamiento se ampara en la finalidad de asistencia sanitaria.

- **Artículo 9:**

Categorías especiales de datos. Regula una protección reforzada de los datos de salud. Obliga a que el entrenamiento y la validación con datos clínicos personales solo sea posible bajo excepciones como el consentimiento o el interés público para la salud o la investigación. Los datos de prueba médicos solo pueden usarse bajo excepciones estrictas (interés público en salud, investigación, consentimiento explícito). El tratamiento de datos de entrada médicos solo es lícito bajo las excepciones del art. 9.2 (atención sanitaria, interés público, etc.).

El Artículo 9 reconoce que los datos de salud son especialmente sensibles y solo pueden utilizarse para entrenar, validar o usar una IA médica cuando exista una excepción legal clara y reforzada, como el consentimiento explícito o una base vinculada a la asistencia sanitaria o al interés público en salud.

Ejemplo:

Una IA que ayuda a predecir complicaciones en pacientes con insuficiencia cardíaca y utiliza datos clínicos como diagnósticos, analíticas y tratamientos.

Se entrena con datos clínicos reales derivadas de historias clínicas. Para ello es necesario que el tratamiento esté amparado por una de las excepciones del artículo 9.2 (consentimiento explícito del paciente, interés público en el ámbito de la salud e investigación científica conforme a la normativa aplicable). Lo mismo sucederá con la validación y los datos de prueba. En cuanto al uso en la práctica clínica, el tratamiento se ampara normalmente en la atención sanitaria o en el interés público.

- **Artículo 22:**

  Decisiones automatizadas. Derecho a no ser objeto de decisiones exclusivamente automatizadas. Limita el uso de la IA sin intervención humana significativa. La IA no puede tomar decisiones clínicas finales basadas solo en datos de entrada sin intervención humana.

El Artículo 22 impide que una IA médica tome decisiones clínicas finales de forma totalmente automática, garantizando siempre una intervención humana significativa en decisiones que afectan de forma relevante a la salud del paciente.

Ejemplo:

Una IA calcula automáticamente si un paciente debe ingresar en UCI tras pasar por urgencias. El paciente tiene el derecho a no ser objeto de una decisión basada exclusivamente en un tratamiento automatizado si la decisión le produce efectos significativos. La IA emite una recomendación tras analizar constantes vitales y antecedentes de “riesgo muy alto. Ingreso en UCI recomendado”. Sin embargo, la consecuencia no puede ser automática, sino que debe intervenir un médico que revise el caso, valore el contexto clínico y confirme o modifique la decisión de la IA. Esto es la intervención humana, que debe ser significativa. Para ello el profesional debe comprender la recomendación, poder cuestionarla y tener capacidad real de cambiarla, ya que el paciente no puede quedar sometido a un sistema que, sin supervisión humana, decida su ingreso, el alta, o el tratamiento.

- **Artículo 25:**

Protección de datos desde el diseño. Obliga a integrar la protección de datos desde la fase de entrenamiento. El sistema debe estar diseñado para proteger los datos de entrada desde su recogida.

El Artículo 25 obliga a que la IA médica esté diseñada desde el principio con medidas técnicas y organizativas que protejan los datos personales, tanto en el entrenamiento como en el uso clínico, y no como un añadido posterior.

Ejemplo:

Una IA analiza imágenes de mamografías para detectar el cáncer de mama. La protección de datos deberá estar integrada desde el inicio del desarrollo del sistema hasta el final. Así, en la fase de entrenamiento, cuando se recopilan miles de mamografías para entrenar a la IA las imágenes se anonimizan antes de usarse, se eliminan los datos identificativos innecesarios y se limita el acceso solo al personal autorizado. El diseño técnico incorpora la privacidad, implicando ello usar estrictamente los datos necesarios, separar los datos identificativos de los clínicos y registrar accesos y actividades para el sistema sea trazable. Cuando el hospital introduce una nueva mamografía -datos de entrada- la transmisión deberá estar cifrada, el almacenamiento deberá ser seguro y no se conservarán datos más allá de los necesarios para la finalidad médica.

- **Artículo 32:**

  Seguridad del tratamiento. Introduce la perspectiva técnica y organizativa en la seguridad. Obliga a proteger los *datasets* de entrenamiento y de validación frente a accesos no autorizados Los conjuntos de datos de prueba deben estar protegidos frente a accesos no autorizados. Los datos de entrada deben estar protegidos frente a accesos no autorizados, pérdidas o filtraciones.

El Artículo 32 obliga a que la IA médica no solo sea clínicamente útil, sino que esté respaldada por medidas técnicas y organizativas sólidas que protejan los datos frente a accesos indebidos, pérdidas o brechas de seguridad.

Ejemplo:

Una IA predice reingresos hospitalarios a partir de historiales clínicos. Los miles de historiales utilizados para entrenar a la IA se almacenan en servidores seguros, lo que implica un acceso restringido solo a personal autorizado, el cifrado de la información y el registro de accesos para detectar usos indebidos. Los conjuntos de datos usados para comprobar el rendimiento del modelo también están protegidos por lo que no pueden copiarse libremente ni descargarse sin control y se aplican políticas internas de seguridad y de auditoría. Se protegen los datos de entrada de uso clínico, por lo que cuando un médico introduce datos del paciente la transmisión está cifrada, el sistema previene pérdidas, accesos no autorizados o filtraciones y existen copias de seguridad.

- **Artículo 35:**

  Evaluación de impacto (EIPD). Implica el análisis previo de tratamientos de alto riesgo. Será obligatoria cuando se entrene una IA con datos sanitarios a gran escala y, en general, cuando la IA sanitaria pueda afectar gravemente a pacientes. Los conjuntos de datos de prueba deben estar protegidos frente a accesos no autorizados. El uso sistemático de IA con datos de entrada sanitarios suele exigir una EIPD previa.

El Artículo 35 obliga a analizar y documentar previamente los riesgos cuando una IA sanitaria pueda tener un impacto elevado en los pacientes, garantizando que se identifiquen y mitiguen antes de su uso real.

Ejemplo:

Una IA analiza millones de historias clínicas para predecir el riesgo de suicidio en pacientes atendidos en atención primaria. Con carácter previo al entrenamiento, como se van a usar datos de salud a gran escala (diagnósticos, tratamientos, antecedentes psicológicos), el hospital debe realizar una EIPD para evaluar riesgos como el uso indebido de datos sensibles, posibles errores que afecten gravemente a pacientes y el impacto en derechos como la confidencialidad o la no discriminación. Con relación a las medidas para reducir riesgos se deberán anonimizar o seudonimizar los datos, limitar los accesos y garantizar una supervisión humana obligatoria en la toma de decisiones críticas. Los conjuntos de datos de prueba utilizados para validar el modelo también deberán protegerse frente a accesos no autorizados.

5. LOS DATOS DE SALUD

De acuerdo con el Considerando 35 del RGPD, entre los datos personales relativos a la salud se deben incluir todos los datos relativos al estado de salud del interesado que dan información sobre su estado de salud física o mental pasado, presente o futuro. Se incluye la información sobre la persona física recogida con ocasión de su inscripción a efectos de asistencia sanitaria, o con ocasión de la prestación de tal asistencia, de conformidad con la Directiva 2011/24/UE del Parlamento Europeo y del Consejo; todo número, símbolo o dato asignado a una persona física que la identifique de manera unívoca a efectos sanitarios; la información obtenida de pruebas o exámenes de una parte del cuerpo o de una sustancia corporal, incluida la procedente de datos genéticos y muestras biológicas, y cualquier información relativa, a título de ejemplo, a una enfermedad, una discapacidad, el riesgo de padecer enfermedades, el historial médico, el tratamiento clínico o el estado fisiológico o biomédico del interesado, independientemente de su fuente, por ejemplo un médico u otro profesional sanitario, un hospital, un dispositivo médico, o una prueba diagnóstica *in vitro*.

Conforme al artículo 4.15 del RGPD datos relativos a la salud son datos personales relativos a la salud física o mental de una persona física, incluida la prestación de servicios de atención sanitaria, que revelen información sobre su estado de salud. Además, el punto 13 hace referencia a los datos genéticos como datos personales relativos a las características genéticas heredadas o adquiridas de una persona física que proporcionen una información única sobre la fisiología o la salud de esa persona, obtenidos en particular del análisis de una muestra biológica de tal persona.

De conformidad con la previsión del artículo 3.37 de la Ley de IA relacionado con el artículo 9.1 del RGPD el tratamiento de esta tipología de datos personales van a gozar de una protección más estricta cuando sean objeto de tratamiento por un sistema de IA.

6. LA BASE DE LEGITIMACIÓN DEL TRATAMIENTO DE DATOS PERSONALES DE SALUD POR LA IA

Hemos visto como los datos son necesarios para el funcionamiento de un sistema de IA. A un sistema de *machine learning* supervisado de una IA que detecta la neumonía en radiografías se le proporcionará la imagen de la radiografía y emitirá como información de salida la probabilidad de neumonía. Para ello previamente habrá sido entrenado con miles de radiografías etiquetadas (con o sin neumonía), posteriormente comprobará sus predicciones con los datos etiquetados y ajustará parámetros para reducir errores. En el caso de aprendizaje no supervisado el sistema agrupará pacientes según patrones en datos clínicos en grupos de pacientes similares. En este caso usa datos sin etiquetas y el algoritmo busca patrones y similitudes automáticamente.

En la medida en que todos esos datos se refieran a una persona o puedan ser referidos a ella el modelo de IA trata datos personales y conforme al RGPD será necesario contar con una base jurídica de tratamiento.

Las bases jurídicas para el tratamiento de datos personales, en general, se encuentran establecidas por el artículo 6 del RGPD, pero los datos de salud son una categoría especial de datos por lo que su tratamiento está prohibido por defecto y solo es lícito si existe una base de legitimación específica, ya que su uso y tratamiento podría entrañar importantes riesgos para los derechos y las libertades fundamentales de la persona.

Además de la previsión realizada en general por el artículo 6, el tratamiento de datos de salud necesita una base legal reforzada y específica según la previsión del artículo 9 del RGPD.

Así, con todo, la prohibición general admite excepciones, como el consentimiento explícito del paciente libre, específico e informado; cuando lo establezca el Derecho de la Unión o de los Estados miembros; cuando esté amparado por motivos de interés público como podría ser la supervisión y alerta sanitaria, la prevención o el control de enfermedades transmisibles y otras amenazas graves para la salud; cuando se fundamente en fines relacionados con la sanidad y con la gestión de los servicios de asistencia sanitaria; y con fines de investigación científica o estadísticos.

Una regla especial que se deriva del ámbito del tratamiento bajo sistemas de IA de alto riesgo es que conforme al artículo 10.5 se permite el tratamiento de datos de salud para detectar y corregir sesgos. Esto implica que la Ley de IA amplía las posibilidades de uso de los datos de salud durante el proceso de entrenamiento de los sistemas de IA de alto riesgo, si bien establece las siguientes condiciones:

- Que el tratamiento de otros datos, como los sintéticos o los anonimizados, no permita efectuar de forma efectiva la detección y corrección de sesgos. Los datos sintéticos son datos generados artificialmente mediante algoritmos o modelos que imitan las características y patrones de datos reales, pero que no corresponden directamente a personas reales.

- Que las categorías especiales de datos personales estén sujetas a limitaciones técnicas relativas a la reutilización de los datos personales y a medidas punteras en materia de seguridad y protección de la intimidad, incluida la seudonimización.

- Que las categorías especiales de datos personales estén sujetas a medidas para garantizar que los datos personales tratados estén asegurados, protegidos y sujetos a garantías adecuadas, incluidos controles estrictos y documentación del acceso, a fin de evitar el uso indebido y garantizar que solo las personas autorizadas tengan acceso a dichos datos personales con obligaciones de confidencialidad adecuadas.

- Que las categorías especiales de datos personales no se transmitan ni transfieran a terceros y que estos no puedan acceder de ningún otro modo a ellos.

- Que las categorías especiales de datos personales se eliminen una vez que se haya corregido el sesgo o los datos personales hayan llegado al final de su período de conservación, si esta fecha es anterior.

- Que los registros de las actividades de tratamiento con arreglo a los Reglamentos (UE) 2016/679 y (UE) 2018/1725 y la Directiva (UE) 2016/680 incluyan las razones por las que el tratamiento de categorías especiales de datos personales era estrictamente necesario para


ANALYSIS
FEMALE

detectar y corregir sesgos, y por las que ese objetivo no podía alcanzarse mediante el tratamiento de otros datos.

La norma permite usar datos especialmente sensibles para corregir sesgos en una IA de alto riesgo solo como último recurso, de forma temporal, controlada y con garantías reforzadas, para proteger tanto la igualdad como los derechos fundamentales de las personas.

Ejemplo:

Una IA de alto riesgo ayuda a priorizar paciente para un trasplante renal. Tras su implantación piloto se detecta que el sistema podría estar asignando sistemáticamente menor prioridad a determinados grupos étnicos. Para comprobar si existe ese sesgo y corregirlo, el proveedor necesita analizar datos especialmente sensibles (por ejemplo, origen étnico o determinadas condiciones de salud). La norma permite tratar excepcionalmente categorías sensibles de datos solo si es estrictamente necesario para detectar y corregir sesgos bajo estrictas condiciones.

Así la necesidad deberá ser real y justificada lo que implica analizar en primer lugar si el sesgo puede detectarse con datos anonimizados o sintéticos. Si no fuera posible, se permite utilizar temporalmente datos sensibles.

Con todo, se tratarán con limitaciones técnicas y de seguridad reforzada como la seudonimización, medidas avanzadas de seguridad y sin posibilidad de reutilizarlos para otros fines.

El acceso será estrictamente controlado de modo que solo el personal autorizado podrá acceder a la información, se registrarán todos los accesos y se establecerán obligaciones estrictas de confidencialidad.

Los datos no se compartirán ni se podrán ceder a terceros y cuando el sesgo haya sido analizado y corregido, los datos sensibles se eliminarán si ya no son necesarios.

Todo deberá quedar registrado.

7. EL PERFILADO Y LA TOMA DE DECISIONES AUTOMATIZADAS

El artículo 22.1 del RGPD garantiza el derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en el interesado o que le afecte significativamente de modo similar.

Pensemos en un sistema de IA de triaje hospitalario que decide si un paciente puede acceder o no a una prueba de diagnóstico urgente. Este sistema requerirá la introducción de los datos de entrada en un formulario digital (dolor, fiebre, edad, antecedentes, entre otros). La IA analizará esos datos y decidirá automáticamente "No cumple criterios: prueba denegada" o "Cumple criterios: prueba autorizada". Todo ello sin intervención de ningún profesional sanitario.

Resulta de aplicación el artículo 22.1 del RGPD porque nos encontramos ante una decisión totalmente automatizada que produce un efecto significativo en el paciente al condicionar el acceso a una prestación sanitaria y al poder retrasar esta decisión el diagnóstico o el tratamiento. El paciente tendría derecho a no ser objeto de esa decisión sin que haya mediado intervención humana.

No obstante, el punto dos del citado precepto nos dice que lo anterior no se aplicará si la decisión:

a. Es necesaria para la celebración o la ejecución de un contrato entre el interesado y un responsable del tratamiento

b. Está autorizada por el Derecho de la Unión o de los Estados miembros que se aplique al responsable del tratamiento y que establezca asimismo medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado

c. Se basa en el consentimiento explícito del interesado

Con todo, en los casos a) y c) el punto 3 añade que el responsable del tratamiento adoptará las medidas adecuadas para salvaguardar los derechos y

libertades y los intereses legítimos del interesado, como mínimo el derecho a obtener intervención humana por parte del responsable, a expresar su punto de vista y a impugnar la decisión.

Si retomamos el ejemplo anterior, una vez la IA emite la recomendación, el médico revisa la recomendación, los datos del paciente y la confirma, la modifica o bien la rechaza, por lo que la decisión final de autorizar o no la prueba la toma el profesional, no la IA. En este caso la decisión ya no es exclusivamente automatizada y no se aplica el artículo 22.1 del RGPD porque existe una intervención humana real y significativa, que debe ser efectiva, es decir, no meramente simbólica -que implicaría aceptar la recomendación de la IA sin analizar y adoptar una decisión humana-.

Además, en lo referente al tratamiento de datos personales de salud, el artículo 22 del RGPD en su punto 4 añade que las decisiones a que se refiere el apartado 2 no se basarán en las categorías especiales de datos personales contempladas en el artículo 9, apartado 1, salvo que se aplique el artículo 9, apartado 2, letra a) o g), y se hayan tomado medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado.

8. LA SUPERVISIÓN HUMANA EN LA IA

La Ley de IA establece requerimientos para garantizar la intervención humana significativa en estos casos en lo que respecta a los sistemas de IA de alto riesgo.

El artículo 14 de la Ley de IA establece la obligación de que los sistemas de IA de alto riesgo se diseñen y se desarrollen de modo que puedan ser vigilados de modo efectivo por personas físicas durante el periodo en el que estén en uso. Ello incluye dotarlos de herramientas de interfaz humano-máquina adecuadas.

Ejemplo:

Un sistema de IA prioriza radiografías de tórax en un hospital para detectar posibles casos graves (por ejemplo, neumonía). El radiólogo tiene una interfaz humano-máquina que muestra: el nivel de riesgo asignado por la

IA (alto-medio-bajo), las zonas de la imagen que han influido en la predicción (mapa de calor) y el grado de confianza del sistema. Con estos elementos el médico puede confirmar la prioridad, cambiarla manualmente o ignorar la recomendación de la IA. El sistema de IA permite pausar o desactivar la IA en cualquier momento y registrar las decisiones humanas para auditorías. En este caso la vigilancia humana es efectiva porque el médico entiende la recomendación, puede intervenir antes de que produzca efectos clínicos y tiene el control real sobre el resultado.

Conforme al artículo 14.2 de la Ley de IA el objetivo de la supervisión humana es prevenir o reducir al mínimo los riesgos para la salud que pueden producirse cuando se utiliza un sistema de IA de alto riesgo conforme a la finalidad prevista o cuando se le da un uso indebido razonablemente previsible.

a. Las medidas de supervisión, conforme al 14.3 deberán ser proporcionales a los riesgos, al nivel de autonomía de la IA y al contexto de uso del sistema y se garantizarán mediante los siguientes tipos de medidas:

Las medidas que el proveedor defina y que integre, cuando sea técnicamente viable, en el sistema de IA de alto riesgo antes de su introducción en el mercado o su puesta en servicio

Ejemplo:

Un sistema de IA recomienda la dosis de insulina para pacientes hospitalizados con diabetes. Antes de su puesta en servicio el proveedor ha introducido medidas de supervisión humana. Ello implica:

- Una interfaz humano-máquina explicativa: muestra la dosis recomendada, los datos usados (glucemia, peso, ingesta reciente) y el nivel de confianza e incluye alertas cuando la recomendación se basa en datos incompletos o atípicos.

- Confirmación humana obligatoria: la dosis no se administra automáticamente y requiere validación explícita del profesional sanitario responsable.

- Capacidad de anulación y ajuste: el profesional puede modificar o rechazar la recomendación y el sistema registra la decisión humana y el motivo

- Límites y salvaguardas predefinidos: el proveedor fija rangos máximos y mínimos de dosis y si la recomendación sale de estos rangos, la IA se bloquea y solicita revisión médica.

- Función de parada o desactivación: botón de apagado inmediato del sistema en caso de comportamiento anómalo.

El sistema, en definitiva, contempla medidas definidas e integradas por el proveedor antes de su comercialización que permiten una supervisión humana efectiva, técnicamente viable y documentada.

b. Las medidas que el proveedor defina antes de la introducción del sistema de IA de alto riesgo en el mercado o de su puesta en servicio y que sean adecuadas para que las ponga en práctica el responsable del despliegue.

Ejemplo:

Sistema de IA que prioriza pacientes en urgencias según riesgo clínico. Se trata en este caso del establecimiento de medidas definidas por el proveedor para aplicar antes del uso. Ello implica:

- Procedimientos de supervisión humana: instrucciones claras para que un médico de triaje revise siempre la clasificación de riesgo generada por la IA antes de que tenga efectos asistenciales.

- Definición de perfiles y de responsabilidades: el proveedor especifica qué profesionales sanitarios están autorizados a supervisar, confirmar o anular decisiones de la IA

- Formación obligatoria: material y requisitos de formación para el personal (funcionamiento del sistema, límites, sesgos y uso correcto)

- Protocolos de actuación ante fallos: guías para identificar resultados anómalos y desactivar el sistema para pasar a un procedimiento manual.

- Uso conforme a la finalidad prevista: instrucciones que prohíben usar la IA para decisiones finales automáticas como puede ser el alta médica si no está diseñada para ello.

En definitiva, estas medidas no se integran técnicamente en el sistema, sino que el proveedor las define y las documenta para que el responsable del despliegue, en este caso el hospital, las establezca en su organización para garantizar la supervisión humana efectiva.

Además, según el punto 4 del citado artículo, los profesionales sanitarios a los que se les encomiende la supervisión deberán:

- Entender adecuadamente las capacidades y limitaciones pertinentes del sistema de IA de alto riesgo y poder vigilar debidamente su funcionamiento, por ejemplo, con vistas a detectar y resolver anomalías, problemas de funcionamiento y comportamientos inesperados.

- Ser conscientes de la posible tendencia a confiar automáticamente o en exceso en los resultados de salida generados por un sistema de IA de alto riesgo («sesgo de automatización»), en particular con aquellos sistemas que se utilizan para aportar información o recomendaciones con el fin de que personas físicas adopten una decisión.

- Interpretar correctamente los resultados de salida del sistema de IA de alto riesgo, teniendo en cuenta, por ejemplo, los métodos y herramientas de interpretación disponibles.

- Decidir, en cualquier situación concreta, no utilizar el sistema de IA de alto riesgo o descartar, invalidar o revertir los resultados de salida que este genere.

- Intervenir en el funcionamiento del sistema de IA de alto riesgo o interrumpir el sistema pulsando un botón de parada o mediante un procedimiento similar que permita que el sistema se detenga de forma segura.

Ejemplo:

Un sistema de IA apoya la decisión médica para la detección de ictus en TAC cerebral en un servicio de urgencias. Los profesionales sanitarios encargados de la supervisión deberán:

- Comprender las capacidades y las limitaciones: el neurorradiólogo sabe que la IA detecta patrones compatibles con ictus isquémico temprano, no sustituye el diagnóstico clínico y puede fallar en imágenes de pacientes con cirugías previas. Esto le permite vigilar el funcionamiento de la IA y detectar resultados anómalos.

- Ser conscientes del sesgo de automatización: el profesional recibe formación específica para no aceptar automáticamente una alerta de ictus probable y tiene que contrastar siempre la salida de la IA con la imagen original y la clínica del paciente.

- Interpretar correctamente los resultados: la interfaz muestra la probabilidad estimada y los mapas de calor que señalan las zonas relevantes. El médico interpreta la salida teniendo en cuenta estas herramientas, no solo el resultado numérico.

- Ser capaz de descartar o no usar el sistema: en un caso concreto (imagen de baja calidad), el profesional decide ignorar la recomendación de la IA o basar la decisión únicamente en su juicio clínico.

- Intervenir y realizar parada segura: el sistema dispone de un botón de parada que permite desactivar la IA si se detecta un comportamiento erróneo y de un procedimiento para volver inmediatamente al flujo manual de trabajo.

En definitiva, el sistema permite una supervisión humana real y efectiva, conforme a las exigencias de la Ley de IA, manteniendo el control clínico en manos del profesional sanitario.

9. OTRAS OBLIGACIONES PARA SISTEMAS DE ALTO RIESGO EN PDP

Lo anterior está en estrecha relación con las obligaciones establecidas en el RGPD, en concreto en el artículo 25, con relación a la protección de datos desde el diseño y por defecto. Ello implica que conforme al estado de la técnica, el coste de la aplicación y la naturaleza, ámbito, contexto y fines del tratamiento, así como los riesgos de diversa probabilidad y gravedad que entraña el tratamiento para los derechos y libertades de las personas físicas, el responsable del tratamiento aplicará, tanto en el momento de determinar los medios de tratamiento como en el momento del propio tratamiento, medidas técnicas y organizativas apropiadas, como la seudonimización.

Ejemplo:

Un sistema de IA que predice el riesgo de reingreso hospitalario a 30 días usando historias clínicas electrónicas. El responsable del tratamiento aplica medidas técnicas y organizativas adecuadas. Ello implica:

- Determinación de los medios de tratamiento (antes del uso): se seleccionan solo los datos necesarios para el modelo (edad, diagnósticos relevantes, número de ingresos previos), excluyendo datos no pertinentes (p. ej., notas clínicas libres no necesarias). Los datos se seudonimizan antes de usarse para entrenar y validar el sistema y se evalúan riesgos mediante una EIPD dada la sensibilidad de los datos de salud.

- Durante el tratamiento: acceso restringido a los datos y a las salidas del sistema solo a personal autorizado, registros de acceso y de uso del sistema (logs) para detectar usos indebidos y separación entre datos identificativos y datos clínicos utilizados por la IA.

- Medidas proporcionales al riesgo: dado que el sistema puede afectar a decisiones asistenciales, se aplican medidas reforzadas de seguridad (cifrado, control de accesos) y se evita el uso de decisiones totalmente automatizadas con efectos jurídicos.

- Integración de garantías: posibilidad de intervención humana y revisión de resultados, así como información clara a los pacientes sobre el uso de IA y sus derechos.

En este caso se trata de un sistema que está diseñado y operado conforme a protección de datos desde el diseño y por defecto, equilibrando estado de la técnica, costes y riesgos, y garantizando los derechos de los interesados.

Además, el responsable del tratamiento aplicará las medidas técnicas y organizativas apropiadas con el objeto de garantizar que, por defecto, solo se traten los datos personales que sean necesarios para cada uno de los fines específicos del tratamiento. Esta obligación se aplicará a la cantidad de datos personales recogidos, a la extensión de su tratamiento, a su plazo de conservación y a su accesibilidad. Tales medidas garantizarán en particular que, por defecto, los datos personales no sean accesibles sin la intervención de la persona a un número indeterminado de personas físicas

Ejemplo:

Sistema de IA que predice el riesgo de sepsis en pacientes hospitalizados a partir de datos clínicos. El responsable del tratamiento aplica la protección de datos por defecto prevista en el artículo 25.2. RGPD:

- Cantidad de datos personales: el sistema solo utiliza los datos estrictamente necesarios para la predicción (signos vitales, analíticas clave). No incorpora datos irrelevantes para la finalidad (p. ej. datos administrativos o sociales).

- Extensión del tratamiento: los datos se usan exclusivamente para la predicción del riesgo de sepsis. Está prohibido su uso para otras finalidades (investigación, gestión) sin base legal adicional.

- Plazo de conservación: los datos utilizados por la IA se conservan solo durante el ingreso hospitalario y el tiempo necesario para auditoría clínica. Posteriormente se eliminan o se anonimizan.

- Accesibilidad: solo el equipo asistencial directamente implicado puede acceder a los datos y a las salidas del sistema. No son visibles para otros servicios ni para personal administrativo.

- No accesibilidad por defecto: las salidas del sistema no se publican ni comparten automáticamente. Cualquier acceso adicional requiere intervención humana autorizada (perfiles, permisos).

En definitiva, este sistema garantiza que, por defecto, el tratamiento de datos personales se limita a lo necesario y que los datos no son accesibles a un número indeterminado de personas, cumpliendo el principio de minimización y privacidad por defecto previsto en el RGPD.

Con relación a la seguridad de los datos personales, el artículo 32 del RGPD establece que teniendo en cuenta el estado de la técnica, los costes de aplicación, y la naturaleza, el alcance, el contexto y los fines del tratamiento, así como riesgos de probabilidad y gravedad variables para los derechos y libertades de las personas físicas, el responsable y el encargado del tratamiento aplicarán medidas técnicas y organizativas apropiadas para garantizar un nivel de seguridad adecuado al riesgo, que en su caso incluya, entre otras:

a. La seudonimización y el cifrado de datos personales.

b. La capacidad de garantizar la confidencialidad, integridad, disponibilidad y resiliencia permanentes de los sistemas y servicios de tratamiento.

c. La capacidad de restaurar la disponibilidad y el acceso a los datos personales de forma rápida en caso de incidente físico o técnico.

d. Un proceso de verificación, evaluación y valoración regulares de la eficacia de las medidas técnicas y organizativas para garantizar la seguridad del tratamiento.

Ejemplo:

Sistema de IA que analiza historias clínicas electrónicas para predecir complicaciones postoperatorias. El responsable y el encargado del tratamiento deberán adoptar las siguientes medidas técnicas y organizativas:

- **Seudonimización y cifrado:** los datos utilizados por la IA se **seudonimizan** antes del entrenamiento y la validación, mientras que los datos en reposo y en tránsito están **cifrados** (bases de datos y comunicaciones internas).

- **Confidencialidad, integridad, disponibilidad y resiliencia:** acceso restringido mediante **control de roles** (solo personal clínico autorizado) y sistemas con **copias de seguridad** y arquitecturas redundantes para evitar pérdida de datos y mecanismos que evitan modificaciones no autorizadas de los datos o del modelo.

- **Restauración rápida tras incidentes: p**rocedimientos de *backup* **y recuperación** que permiten restablecer el acceso a los datos clínicos en caso de fallo técnico o ciberataque y plan de continuidad asistencial para que el hospital pueda seguir trabajando sin la IA si es necesario.

- **Evaluación y verificación periódicas: auditorías de seguridad** y pruebas periódicas (p. ej., tests de vulnerabilidad) y revisión regular de permisos y registros de acceso y actualización de medidas conforme evoluciona el **estado de la técnica**.

Así, el sistema de IA médica garantizará un nivel de seguridad adecuado al riesgo, protegiendo datos de salud y los derechos de los pacientes conforme al artículo 32 del RGPD.

10. MEDIDAS DE GOBERNANZA DE DATOS EN LA LEY DE IA Y SESGOS DE LA IA

También resulta de interés en cuanto a las obligaciones adicionales para los sistemas de IA de alto riesgo las previsiones del artículo 10 de la Ley de IA con relación a las medidas a adoptar por el proveedor de un sistema de IA de alto riesgo relacionadas con la gobernanza de los datos y de aplicación a los conjuntos de datos de entrenamiento, validación y prueba. En concreto, las prácticas de gobernanza se deberán centrar en:

- Las decisiones pertinentes relativas al diseño.
- Los procesos de recogida de datos y el origen de los datos y, en el caso de los datos personales, la finalidad original de la recogida de datos.
- Las operaciones de tratamiento oportunas para la preparación de los datos, como la anotación, el etiquetado, la depuración, la actualización, el enriquecimiento y la agregación.
- La formulación de supuestos, en particular en lo que respecta a la información que miden y representan los datos.
- Una evaluación de la disponibilidad, la cantidad y la adecuación de los conjuntos de datos necesarios.
- El examen atendiendo a posibles sesgos que puedan afectar a la salud y a la seguridad de las personas, afectar negativamente a los derechos fundamentales o dar lugar a algún tipo de discriminación prohibida por el Derecho de la Unión, especialmente cuando las salidas de datos influyan en las informaciones de entrada de futuras operaciones.
- Medidas adecuadas para detectar, prevenir y mitigar posibles sesgos detectados.
- La detección de lagunas o deficiencias pertinentes en los datos que impidan el cumplimiento de la Ley de IA, y la forma de subsanarlas.

Especial interés en esta materia tienen los sesgos o errores en los datos para el entrenamiento, validación y prueba.

Los sesgos en los datos de salud aparecen cuando la información con la que se entrena o se evalúa una IA no representa bien a toda la población o contiene errores sistemáticos. Ya hemos tenido ocasión de comprobar que la IA aprende de los datos que se le proporcionan y si los datos están desequilibrados o incompletos, el resultado de salida de la IA también lo estará.

Si una IA que detecta cáncer de piel es entrenada casi solo con imágenes de personas de piel clara funcionará bien en ese grupo, pero fallará más en personas con piel oscura, porque los datos adolecen de calidad y de representatividad.

Los sesgos pueden provocar más errores en ciertos grupos (edad, sexo, origen étnico), infradiagnósticos o sobrediagnósticos en poblaciones concretas y desigualdades en la atención sanitaria.

La IA no es neutral por sí misma y refleja los patrones y desequilibrios presentes en los datos por lo que se puede afirmar que “basura que entra, basura que sale”. Es por ello, que resulta fundamental usar datos diversos y representativos, detectar y corregir sesgos antes del uso clínico y mantener la supervisión humana.

Así, los sesgos en datos de salud pueden traducirse en decisiones clínicas menos justas o seguras, y por eso su identificación y corrección es esencial en cualquier sistema de IA sanitaria.

Ejemplo:

Sistema de IA que predice el riesgo cardiovascular y recomienda intensificar tratamiento preventivo.

Examen de posibles sesgos (antes y durante el uso) El proveedor y el hospital analizan sesgos específicos de la IA médica que pueden afectar a la salud, la seguridad y los derechos fundamentales:

- Sesgo de datos: el conjunto de entrenamiento contiene más datos de hombres que, de mujeres, y más de población adulta media frente a

personas mayores. El riesgo radica en infradiagnóstico en mujeres o ancianos.

- Sesgo histórico o clínico: los datos reflejan prácticas médicas pasadas donde ciertos grupos (p. ej. mujeres) recibían menos pruebas diagnósticas. La IA podría reproducir desigualdades existentes.

- Sesgo de medición: variables clínicas (p. ej., dolor torácico) se registran de forma distinta según el sexo o el origen cultural del paciente.

- Sesgo de retroalimentación (*feedback loop*): si la IA recomienda menos tratamiento preventivo a un grupo, en el futuro habrá menos datos clínicos de ese grupo, reforzando el sesgo en nuevas versiones del modelo.

Medidas para detectar, prevenir y mitigar sesgos:

- Análisis de rendimiento por subgrupos: evaluación separada por sexo, edad, origen étnico y comorbilidades, además de comparación de tasas de falsos negativos y falsos positivos.

- Curación y balanceo de datos: inclusión deliberada de datos infrarrepresentados y revisión clínica de etiquetas para evitar errores sistemáticos.

- Controles en el diseño del sistema: límites que impiden decisiones automáticas sin revisión humana en grupos de riesgo y alertas cuando la IA se usa fuera de la población para la que fue validada.

- Supervisión humana reforzada: formación específica a profesionales sobre sesgo de automatización y sesgos algorítmicos y revisión clínica obligatoria de recomendaciones en colectivos vulnerables.

- Monitorización continua: detección de cambios en el rendimiento del modelo tras su despliegue y reentrenamiento controlado cuando se detectan desviaciones injustificadas.

El sistema incorpora un examen sistemático de sesgos y medidas activas de mitigación, reduciendo riesgos para la salud, evitando discriminaciones

prohibidas y cumpliendo las exigencias de la Ley de IA, especialmente críticas en el ámbito de la IA médica.

Además, el artículo 10 añade que "Los conjuntos de datos de entrenamiento, validación y prueba serán pertinentes, suficientemente representativos y, en la mayor medida posible, carecerán de errores y estarán completos en vista de su finalidad prevista. Asimismo, tendrán las propiedades estadísticas adecuadas, por ejemplo, cuando proceda, en lo que respecta a las personas o los colectivos de personas en relación con los cuales está previsto que se utilice el sistema de IA de alto riesgo. Los conjuntos de datos podrán reunir esas características para cada conjunto de datos individualmente o para una combinación de estos".

Ejemplo:

Sistema de IA que detecta cáncer de piel a partir de imágenes dermatológicas.

Pertinencia y finalidad:

- Los datos utilizados son imágenes clínicas de lesiones cutáneas, tomadas con dispositivos similares a los del entorno real de uso.

- Se excluyen imágenes cosméticas o no diagnósticas, por no ser pertinentes para la finalidad médica.

Representatividad de los conjuntos de datos:

- El conjunto total (entrenamiento + validación + prueba) incluye: distintos fototipos de piel (I–VI), diversidad de edad y sexo, lesiones benignas y malignas en proporciones clínicamente realistas.

- Se evita un sesgo frecuente en IA médica: bajo rendimiento en pieles oscuras por infrarrepresentación en los datos

Ausencia de errores y datos completos:

- Las etiquetas ("melanoma", "lesión benigna") se validan mediante: diagnóstico histopatológico o consenso de varios dermatólogos.

- Se eliminan imágenes duplicadas, borrosas o mal etiquetadas.

- Los metadatos esenciales (localización, edad aproximada) están completos.

Propiedades estadísticas adecuadas:

- Se comprueba que la distribución de casos en los datos se corresponde con la población prevista de uso (atención primaria y dermatología hospitalaria) y mantiene proporciones similares entre conjuntos de entrenamiento, validación y prueba.

- El rendimiento se evalúa por subgrupos (fototipo, edad), no solo de forma global.

Uso combinado de conjuntos de datos:

- Si un único conjunto no es suficientemente representativo se combinan datos de varios hospitales y países, manteniendo coherencia clínica y controles de calidad.

- Cada conjunto individual puede cubrir una parte, pero la combinación final cumple los requisitos del artículo 10.

El sistema se entrena y evalúa con datos pertinentes, representativos, completos y estadísticamente adecuados, reduciendo sesgos clínicos y riesgos para la salud, conforme al artículo 10 de la Ley de IA.

Por último, los conjuntos de datos tendrán en cuenta, en la medida necesaria para la finalidad prevista, las características o elementos particulares del entorno geográfico, contextual, conductual o funcional específico en el que está previsto que se utilice el sistema de IA de alto riesgo.

Ejemplo:

IA que predice el riesgo de sepsis en servicios de urgencias hospitalarias.

Consideración del entorno geográfico:

- Los datos incluyen pacientes de hospitales de la región donde se desplegará la IA.
- Se tienen en cuenta la prevalencia local de infecciones y los patrones de resistencia antimicrobiana.
- Se evita un problema común en IA médica, que proviene de modelos entrenados en otros países que rinden peor en contextos locales.

Consideración del contexto asistencial:

- Los datos reflejan el flujo real de urgencias: tiempos de espera, frecuencia de analíticas y disponibilidad de recursos.
- El modelo se entrena para funcionar con datos incompletos, habituales en urgencias.

Consideración de factores conductuales:

- Se incorporan patrones reales de adherencia a protocolos clínicos y frecuencia de toma de constantes.
- Se evita entrenar con datos “ideales” que no existen en la práctica diaria.

Consideración del entorno funcional:

- Los datos proceden de los mismos sistemas de información clínica y formatos que se usarán en producción.
- Se tienen en cuenta diferencias entre hospitales grandes y comarcales, y entre turnos diurnos y nocturnos.

Los conjuntos de datos están adaptados al entorno geográfico, contextual, conductual y funcional real, garantizando que la IA médica de alto riesgo funcione de forma segura, fiable y conforme a su finalidad prevista, tal como exige la Ley de IA.

Un ejemplo paradigmático en esta materia es el "Síndrome de Yentl", que describe cómo las mujeres pueden recibir un diagnóstico o tratamiento adecuado solo cuando presentan síntomas típicamente masculinos. El término procede de un artículo publicado en 1991 en *The British Medical Journal* por la cardióloga estadounidense Bernardine Healy, inspirándose en el personaje de ficción de Yentl (una mujer que debía hacerse pasar por hombre para acceder a ciertos derechos).

Durante mucho tiempo, las investigaciones médicas -especialmente en enfermedad cardiovascular- se basaron principalmente en datos de hombres. La consecuencia es que los síntomas "clásicos" del infarto como dolor intenso en el pecho irradiado al brazo izquierdo, se definieron a partir de patrones masculinos. Al presentar las mujeres síntomas diferentes como fatiga, náuseas, dolor en la espalda o en la mandíbula, no eran diagnosticadas a tiempo.

Se trata de un sesgo que se ha ido repitiendo y que afecta a la alimentación de la IA por lo que debe corregirse ya que de lo contrario las mujeres son infra-diagnosticadas e infratratadas.

11. DERECHOS SOBRE DATOS PERSONALES EN IA

La normativa europea en materia de protección de datos contempla derechos para los interesados cuyos datos personales son contemplados en los artículos 12 a 22 del RGPD.

En particular, los interesados tienen derecho de:

a. **Acceso (artículo 15):** conocer si sus datos están siendo tratados por un sistema de IA, qué datos concretos se usan, con qué finalidad y cómo influyen en el funcionamiento del sistema.

Ejemplo:

Sistema de IA que **predice el riesgo de reingreso hospitalario** usando la historia clínica electrónica. Un paciente ejerce su **derecho de acceso** y solicita información sobre el uso de sus datos por el sistema de IA. El responsable del tratamiento debe facilitarle:

- **Confirmación** de que sus datos personales han sido tratados por un sistema de IA.
- **Qué datos concretos** se han utilizado: edad, diagnósticos previos, número de ingresos anteriores, resultados analíticos relevantes.
- **Finalidad del tratamiento**: apoyar a los profesionales sanitarios en la planificación del alta y el seguimiento.
- **Lógica general del sistema** (a nivel comprensible): que la IA identifica patrones estadísticos en datos clínicos para estimar el riesgo, sin revelar secretos comerciales ni el código fuente.
- **Resultados que le conciernen**: por ejemplo, que fue clasificado como "riesgo medio de reingreso".
- **Destinatarios** de los datos: equipo médico responsable y sistema hospitalario.
- **Plazo de conservación** de los datos usados por la IA.
- **Información sobre sus derechos** (rectificación, limitación, oposición, etc.).

El derecho de acceso **no obliga a explicar el algoritmo con detalle**, pero sí a permitir que el paciente **comprenda cómo y por qué sus datos personales han sido utilizados por la IA** y qué efectos puede tener sobre su atención sanitaria.

b. Rectificación (artículo 16): solicitar la corrección de datos inexactos o incompletos, evitando que la IA genere resultados erróneos basados en información incorrecta.

Ejemplo:

Sistema de IA que **predice el riesgo de reingreso hospitalario** a partir de la historia clínica electrónica. Un paciente revisa la información facilitada tras ejercer su derecho de acceso y detecta un **dato personal inexacto** utilizado por la IA (por ejemplo, un diagnóstico previo que no le corresponde).

- El paciente solicita la **corrección del dato erróneo** (p. ej., se le atribuye una insuficiencia cardíaca inexistente).
- El hospital verifica la solicitud mediante revisión clínica.
- El dato incorrecto se **corrige en la historia clínica** y en los sistemas conectados a la IA.
- La IA **deja de utilizar el dato inexacto** en futuras predicciones sobre ese paciente.
- Si el resultado previo (riesgo de reingreso) se vio afectado, se recalcula la predicción, o se deja constancia de que el resultado anterior estaba basado en datos incorrectos.

El derecho de rectificación garantiza que la IA no genere resultados clínicos basados en información errónea, protegiendo tanto los derechos del paciente como la seguridad asistencial.

c. Supresión-Derecho al olvido (artículo 17): pedir la eliminación de sus datos cuando ya no sean necesarios, el tratamiento sea ilícito o se retire el consentimiento, salvo excepciones legales (p. ej. obligaciones en materia de documentación clínica derivadas de las obligaciones de

conservación de los datos sanitarios procedentes de la Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica).

Ejemplo:

Sistema de IA que predice el riesgo de reingreso hospitalario usando datos de la historia clínica electrónica. Un paciente solicita la supresión de sus datos personales utilizados por el sistema de IA tras haber finalizado su tratamiento (supeditado a la conservación de determinada información conforme a las previsiones establecidas por la Ley 41/2002).

- **Análisis previo del responsable del tratamiento:** el hospital comprueba si los datos **ya no son necesarios** para la finalidad asistencial y que no están sujetos a **obligaciones legales de conservación** y no se requieren para la **defensa de reclamaciones**.

- **Ejercicio del derecho de supresión:** los datos personales del paciente se eliminan de los conjuntos operativos usados por la IA, o se **anonimizan de forma irreversible** cuando la supresión total no es posible. Se impide que el sistema de IA vuelva a utilizar esos datos para nuevas predicciones y reentrenamientos futuros.

- **Límites en el ámbito sanitario:** el derecho puede **no aplicarse plenamente** si la conservación es necesaria para fines de salud pública, investigación científica con garantías o cumplimiento de obligaciones legales.

El derecho de supresión en IA médica no implica borrar el modelo entrenado, pero sí asegurar que los datos personales identificables del interesado dejen de ser tratados.

d. Limitación del tratamiento (artículo 18): exigir que los datos no se usen temporalmente por la IA mientras se verifica su exactitud o la licitud del tratamiento.

Ejemplo:

Sistema de IA que predice el riesgo de reingreso hospitalario a partir de la historia clínica electrónica. Un paciente considera que algunos datos usados por la IA pueden ser inexactos y solicita la limitación del tratamiento mientras se verifica su corrección.

- El hospital **marca los datos del paciente como "limitados"** en los sistemas conectados a la IA.
- Durante ese periodo la IA **no utiliza esos datos** para nuevas predicciones, ni para reentrenamiento del modelo.
- Los datos solo se conservan para su verificación, el cumplimiento de obligaciones legales o la defensa de posibles reclamaciones.
- Las salidas de la IA que afecten al paciente se suspenden o se realizan sin usar los datos limitados.

El equipo médico es informado de que **los resultados automáticos pueden no estar disponibles** temporalmente.

La limitación del tratamiento actúa como una **medida de cautela**, evitando que la IA genere decisiones o recomendaciones basadas en datos cuya exactitud o licitud está en discusión.

e. Oposición (artículo 21): oponerse al tratamiento de sus datos en determinadas circunstancias, especialmente cuando la IA se utilice con fines distintos de la asistencia sanitaria directa.

Ejemplo:

Sistema de IA que **predice el riesgo de reingreso hospitalario** utilizando datos de la historia clínica electrónica. El hospital utiliza la IA no solo para la asistencia directa, sino también para **mejorar la gestión de recursos** (p. ej., planificación de camas), basándose en interés público o interés legítimo.

- El paciente se **opone al uso de sus datos** para esa finalidad secundaria.
- El responsable del tratamiento analiza si existen **motivos legítimos imperiosos** que prevalezcan sobre los derechos del interesado.

Si la oposición es estimada:

- Los datos del paciente **dejan de utilizarse** para la finalidad de gestión o planificación.
- Pueden seguir tratándose para la **asistencia sanitaria directa** cuando exista base legal.
- El sistema de IA incorpora una **restricción por finalidad**, respetando la oposición.

Hay que considerar los límites de aplicación de este derecho en el ámbito sanitario, ya que el derecho de oposición **no suele aplicarse** cuando el tratamiento es necesario para la prestación de asistencia sanitaria, o por razones de interés público esencial en salud.

Con todo, el derecho de oposición permite al interesado **controlar usos no estrictamente asistenciales de sus datos por sistemas de IA**, evitando tratamientos adicionales que no considere justificados.

En el contexto de la IA, estos derechos son esenciales para **evitar decisiones injustas, errores clínicos o usos indebidos de los datos**, y se conectan directamente con principios como la **transparencia**, la **intervención humana** y la **responsabilidad** exigidos tanto por el RGPD como por la Ley de IA.

12. UN HITO PARA LOS DATOS Y LA IA: EL ESPACIO EUROPEO DE DATOS DE SALUD.

La aplicación del Reglamento (UE) 2025/327 del Parlamento Europeo y del Consejo, de 11 de febrero de 2025, relativo al Espacio Europeo de Datos de Salud, y por el que se modifican la Directiva 2011/24/UE y el Reglamento (UE) 2024/2847 implica un cambio de paradigma en el uso de los datos personales de salud.

La creación del Espacio Europeo de Datos de Salud (en adelante, EEDS) persigue la mejora del acceso por parte de las personas físicas a sus datos sanitarios electrónicos personales, así como su control por las mismas, tanto en el contexto asistencial -uso primario- como para otros fines en beneficio de la sociedad, como la investigación, la innovación, la formulación de políticas, la seguridad de los pacientes, la medicina personalizada, las estadísticas o la reglamentación -uso secundario-. Además, la Propuesta se dirige a la mejora del mercado interior mediante el establecimiento de un marco jurídico uniforme en lo relativo al desarrollo, comercialización y uso de los sistemas de historiales médicos electrónicos -en adelante, HME- conforme a los valores de la Unión.

El EEDS implica superar el uso de los datos primarios -con fines asistenciales- y trasciende hacia un uso secundario -con fines de investigación, de innovación, reguladores o industria-. Por lo tanto, su materialización implicará avances en la asistencia sanitaria -uso primario- pero también en la investigación sanitaria, en la innovación, en la formulación de políticas y en la regulación -uso secundario-.

Así, la propuesta apunta que el EEDS promoverá el intercambio y el acceso a los diferentes tipos de datos sanitarios electrónicos, incluyendo el HME, los datos genómicos, los registros de pacientes, entre otras fuentes.

Entre los ejemplos concretos que se plantean, la Comisión Europea hace referencia a una empresa de tecnología sanitaria que desarrolla una nueva herramienta de apoyo a la toma de decisiones médicas basadas en la IA que ayuda a los médicos a tomar decisiones diagnósticas y de tratamiento a partir de la revisión de las imágenes de laboratorio del paciente. Por medio de la IA -cuyo soporte para el entrenamiento ha sido el *big data*- es posible comparar las imágenes del paciente con las de muchos pacientes. A través

del EEDS la empresa puede acceder a ingentes cantidades de imágenes médicas que entrenan al algoritmo y optimizan su precisión y eficacia, con carácter previo a solicitar la autorización de comercialización

En su artículo 53 se establece que los organismos de acceso a datos de salud podrán conceder a un usuario de datos de salud acceso a datos de salud electrónicos para uso secundario si el tratamiento de los datos por ese usuario de datos de salud es necesario para (apartado e.) la investigación científica relacionada con el sector sanitario o asistencial que contribuya a la salud pública o a la evaluación de tecnologías sanitarias o que procure niveles elevados de calidad y seguridad de la asistencia sanitaria, de los medicamentos o de los productos sanitarios, con el objetivo de beneficiar a los usuarios finales, como los pacientes, los profesionales sanitarios y los administradores sanitarios, lo que incluye: i) las actividades de desarrollo e innovación para productos o servicios, y ii) el entrenamiento, la prueba y la evaluación de algoritmos, también con respecto a productos sanitarios, productos sanitarios para diagnóstico *in vitro*, sistemas de IA y aplicaciones sanitarias digitales.

Ejemplo:

Necesidad de entrenar un modelo de IA que prediga reingresos hospitalarios.

- Un hospital tiene historiales clínicos electrónicos con fines de atención sanitaria. Si se utilizan para entrenar a un sistema de IA se usarán con fines secundarios.

- Se solicitará a la autoridad nacional de acceso a los datos el uso para el entrenamiento y la evaluación del algoritmo.

- Antes del uso se eliminarán los identificadores directos (como nombre y DNI), los datos se seudonimizan y se limitarán a las variables necesarias.

- Se analizarán en un entorno seguro sin poder descargar los datos que solo se utilizarán para el entrenamiento del modelo

- En el EEDS, los datos clínicos recogidos para tratar pacientes se reutilizan, de forma protegida y controlada, dentro de un entorno seguro para entrenar algoritmos de IA sin exponer información personal.

13. NORMATIVA

Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos).

Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (en adelante, Ley de IA) es el primer marco jurídico global en materia de IA, que aborda sus riesgos con el objetivo de fomentar una IA fiable en Europa.

Reglamento (UE) 2025/327 del Parlamento Europeo y del Consejo, de 11 de febrero de 2025, relativo al Espacio Europeo de Datos de Salud, y por el que se modifican la Directiva 2011/24/UE y el Reglamento (UE) 2024/2847

14. BIBLIOGRAFÍA

1. Gil Membrado, C., "IA y salud cardiovascular en la mujer: una perspectiva biológica y de género", en Gil Membrado, C., Luquin Bergareche, R. (dirs.), *Medicina personalizada genética y criogenización*, Dykinson, 2025.

2. Gil Membrado, C., "La Ley de IA: entre la salud y el Derecho en una nueva era. Un acompasamiento necesario", *Revista Jurídica de las Illes Balears*, núm. 26, 18 de diciembre de 2024.

3. Gil Membrado, C., "Daños producidos por la IA: La opacidad del algoritmo y el efecto caja negra" en *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, Asociación de profesores de Derecho Civil, Aranzadi La Ley, 2024.

4. Gil Membrado, C., "Diabetes y Espacio Europeo de Datos Sanitarios", en Gil Membrado, C., Luquin Bergareche, R. (dirs), *Salud Digital. Aplicaciones Móviles, Telemedicina y Chatbots*, Dykinson, 2024.

5. Gil Membrado, C., *Riesgos del uso de algoritmos en el diagnóstico y en la investigación biomédica*, VIII Premio Nacional de Derecho Sanitario, Reus, 2023.

6. Gil Membrado, C., "En el horizonte: la Directiva de responsabilidad extracontractual en materia de IA", en Gil Membrado, C., Luquin Bergareche, R. (dirs.), *Derecho y medicina: desafíos tecnológicos y científicos*, Dykinson, 2023.

7. Gil Membrado, C. "La apertura del código fuente y la transparencia en la contratación de servicios. Equidad y transparencia en la prestación de servicios", en Cobas Cobiella, M.E., Guillén Catalán, R., (dirs.), *Equidad y transparencia en la prestación de servicios*, Dykinson, 2023.

8. Gil Membrado, C., "Un régimen europeo de responsabilidad civil para el usuario de la inteligencia artificial", *Direito do consumidor no cenário iberoamericano*, Foco, Brasil, 2023.

9. Gil Membrado, C., "Una nueva era: hacia el robot sanitario autónomo y su encaje en el derecho". Bioderecho y retos, en Gil Membrado, C., Luquin Bergareche, R. (dirs.), *Bioderecho y retos. M-Health, genética, robótica y criogenización*, Dykinson, 2022.

10. Gil Membrado, C., "Telemedicina y aplicaciones móviles. la m-Health y la seguridad de los datos personales", en Martínez Valencoso, M.L., Sancho López, M. (dirs.), *Protección jurídica de la privacidad: inteligencia artificial, salud y contratación,* Aranzadi La Ley, 2022.

11. ANEXOS

15.1. CHECK LIST BÁSICO PARA TRATAMIENTO DE DATOS PERSONALES DE SALUD POR SISTEMAS DE IA

1. ¿Se tratan datos personales?

- ☐ ¿Los datos permiten identificar directa o indirectamente a una persona?
- ☐ ¿Existe riesgo razonable de reidentificación, incluso tras anonimización?
- ☐ ¿Se trata de datos de salud (categoría especial del RGPD)?

2. Base jurídica (RGPD arts. 6 y 9)

- ☐ ¿Existe una base legal válida para el tratamiento?
- ☐ ¿Se cumple alguna excepción del art. 9.2 RGPD (consentimiento explícito, asistencia sanitaria, interés público en salud, investigación científica)?
- ☐ ¿Está documentada la base jurídica aplicable en cada fase (entrenamiento, validación, prueba y uso clínico)?

3. Principios del tratamiento (art. 5 RGPD)

- ☐ ¿Los datos son adecuados y limitados a lo estrictamente necesario (minimización)?
- ☐ ¿Son exactos y están actualizados?
- ☐ ¿Se usan solo para la finalidad médica prevista (limitación de finalidad)?
- ☐ ¿Se informa al paciente de forma clara y transparente?

4. Fase de entrenamiento

- ☐ ¿Los datos de entrenamiento son representativos y de calidad?
- ☐ ¿Se han identificado y mitigado posibles sesgos?
- ☐ ¿Se han aplicado medidas de seudonimización o anonimización cuando sea posible?
- ☐ ¿Existe control de accesos y registro de actividades?

5. Fase de validación

- ☐ ¿Se utilizan datos distintos a los de entrenamiento?
- ☐ ¿Se ha evaluado la generalización del modelo?
- ☐ ¿Se han detectado y corregido sobreajustes (*overfitting*)?
- ☐ ¿La validación está documentada técnicamente?

6. Datos de prueba

- ☐ ¿Son independientes del entrenamiento y validación?
- ☐ ¿Permiten evaluar precisión, errores, sesgos y seguridad clínica?
- ☐ ¿Están protegidos frente a accesos no autorizados?

7. Datos de entrada (uso clínico)

- ☐ ¿Los datos introducidos son correctos, completos y actualizados?
- ☐ ¿El sistema detecta datos incompletos o erróneos?
- ☐ ¿Existe supervisión humana significativa?
- ☐ ¿Se evita la toma de decisiones exclusivamente automatizadas?

8. Supervisión humana (Ley de IA art. 14 / RGPD art. 22)

- ☐ ¿El profesional sanitario puede revisar, corregir o ignorar la salida de la IA?
- ☐ ¿La intervención humana es real y no meramente formal?
- ☐ ¿Existen herramientas de interfaz que permitan entender la recomendación de la IA?

9. Gestión de riesgos (Ley de IA art. 9)

- ☐ ¿Se han identificado riesgos clínicos antes del uso?
- ☐ ¿Se han analizado riesgos derivados de datos sesgados o defectuosos?
- ☐ ¿Se gestionan riesgos derivados de datos de entrada incorrectos?
- ☐ ¿Existe sistema de monitorización continua?

10. Gobernanza de datos (Ley de IA art. 10)

- ☐ ¿Los datos son relevantes y representativos para el fin previsto?
- ☐ ¿Existe control sobre la calidad y trazabilidad de los datos?
- ☐ ¿El sistema está preparado para gestionar datos atípicos?

11. Seguridad del tratamiento (art. 32 RGPD)

- ☐ ¿Los *datasets* están cifrados?
- ☐ ¿Hay control estricto de accesos?
- ☐ ¿Se registran accesos y modificaciones?
- ☐ ¿Existen copias de seguridad seguras?

12. Protección de datos desde el diseño (art. 25 RGPD)

- ☐ ¿La privacidad está integrada desde la fase de diseño?
- ☐ ¿Se limita la reutilización de datos?
- ☐ ¿Se separan datos identificativos de datos clínicos?

13. Evaluación de Impacto (art. 35 RGPD)

- ☐ ¿Se ha realizado una EIPD cuando el tratamiento es de alto riesgo?
- ☐ ¿Se han documentado riesgos y medidas mitigadoras?
- ☐ ¿La EIPD cubre todo el ciclo de vida del sistema?

14. Documentación técnica (Ley de IA art. 11)

- ☐ ¿Existe expediente técnico completo?
- ☐ ¿Se documenta el origen y tratamiento de los datos?
- ☐ ¿Se conservan evidencias de validación y pruebas?

15. Conformidad y comercialización (Ley de IA art. 47)

- ☐ ¿Se ha realizado evaluación de conformidad?
- ☐ ¿Se ha emitido declaración formal de conformidad?
- ☐ ¿El rendimiento demostrado es seguro y fiable?

VERIFICACIÓN FINAL

- ☐ ¿El sistema protege los derechos fundamentales del paciente?
- ☐ ¿Existe coherencia entre RGPD y Ley de IA?
- ☐ ¿Se garantiza la seguridad, transparencia y control humano?

15.2. DECÁLOGO DE TRATAMIENTO DE DATOS PERSONALES POR SISTEMAS DE IA DE SALUD

1. La IA médica depende de datos de salud especialmente protegidos

Los datos sanitarios son categorías especiales de datos (RGPD art. 9) y requieren una protección reforzada en todas las fases del ciclo de vida del sistema.

2. No hay IA sanitaria sin base jurídica válida

El tratamiento de datos personales solo es lícito si existe una base legal (art. 6 RGPD) y una excepción específica para datos de salud (art. 9.2 RGPD), como asistencia sanitaria, interés público o consentimiento explícito.

3. Los datos deben ser mínimos, adecuados y exactos

Solo pueden utilizarse los datos estrictamente necesarios para la finalidad clínica prevista (principio de minimización), y deben ser correctos y actualizados.

4. La protección debe integrarse desde el diseño

La privacidad no se añade al final, sino que debe incorporarse desde la fase de desarrollo (art. 25 RGPD), mediante seudonimización, control de accesos y limitación de reutilización.

5. La calidad y representatividad de los datos es esencial

Los datos de entrenamiento, validación y prueba deben ser relevantes, files y representativos para evitar sesgos y errores clínicos (Ley de IA).

6. La IA debe validarse antes de usarse con pacientes

No basta con entrenar el sistema, sino que debe probarse con datos independientes para garantizar precisión, seguridad y generalización antes del uso clínico.

7. La supervisión humana es obligatoria

La IA médica no puede sustituir al profesional sanitario. Debe existir intervención humana significativa, con capacidad real de revisar, corregir o rechazar la salida del sistema.

8. No a las decisiones exclusivamente automatizadas

Cuando la decisión tenga efectos relevantes sobre el paciente, no puede basarse únicamente en un sistema automatizado (art. 22 RGPD), salvo excepciones estrictas y con garantías.

9. Seguridad técnica y organizativa reforzada

Los *datasets* de entrenamiento, validación, prueba y los datos de entrada deben estar protegidos mediante cifrado, control de accesos, trazabilidad y medidas frente a brechas de seguridad (art. 32 RGPD).

10. Evaluación de riesgos y documentación continua

Debe realizarse una evaluación de impacto cuando el tratamiento sea de alto riesgo (art. 35 RGPD) y documentarse todo el ciclo de vida del sistema (Ley de IA), garantizando trazabilidad, control y responsabilidad.

15.3. DECÁLOGO PARA EL MÉDICO: USO RESPONSABLE DE DATOS PERSONALES MEDIANTE SISTEMAS DE IA

1. Recuerda: la IA trabaja con datos de salud

Los datos que introduces (síntomas, pruebas, antecedentes) son datos especialmente protegidos. Su uso debe estar justificado y limitado a la finalidad asistencial.

2. Introduce solo la información necesaria

No añadas datos irrelevantes.

La IA debe trabajar con lo estrictamente necesario para el caso clínico.

3. Verifica la calidad de los datos antes de confiar en el resultado

Si los datos están incompletos o mal registrados, la predicción puede ser errónea.

4. La decisión final es tuya

La IA es herramienta de apoyo.

No sustituye tu juicio clínico ni tu responsabilidad profesional.

5. No aceptes automáticamente la recomendación

Debes poder:

- Revisarla
- Entenderla
- Cuestionarla
- Modificarla

Si no puedes hacerlo, no hay supervisión humana real.

6. Ten en cuenta los límites del sistema

Pregunta o revisa:

- ¿En qué población fue entrenada?
- ¿Qué tasa de errores tiene?
- ¿En qué situaciones puede fallar?

La IA puede no funcionar igual en todos los pacientes.

7. Especial cuidado en decisiones de alto impacto

Ingresos, altas, tratamientos invasivos o denegaciones de pruebas no deben basarse únicamente en la IA.

8. Protege la confidencialidad en el uso diario

Evita:

- Introducir datos en sistemas no autorizados
- Usar herramientas de IA externas sin validación institucional
- Compartir resultados fuera del entorno seguro

9. Participa en la detección de errores y sesgos

Si observas fallos sistemáticos o resultados extraños:

- Comunícalo.
- Documenta el caso.
- No normalices el error.

Tu experiencia es clave para mejorar el sistema.

10. La IA aprende de la práctica clínica

En algunos sistemas, tu validación o corrección puede alimentar el aprendizaje futuro.

Tu criterio ayuda a que el sistema mejore o perpetúe errores si no se revisa críticamente.

El Libro Blanco de IA en Medicina es una obra innovadora y pionera en el ámbito colegial español, llamada a convertirse en un texto de referencia para abordar, desde el rigor, la prudencia y la vocación de servicio, uno de los grandes desafíos de la medicina contemporánea: la integración de la inteligencia artificial en la práctica asistencial, la formación, la organización sanitaria y la relación con los pacientes. Impulsado por el Colegio de Médicos de Málaga a través del Commálaga Health Hub, que dirige y lidera el Dr.José Antonio Trujillo, este volumen representa una apuesta decidida por situar a la profesión médica en el centro del debate sobre innovación y salud digital.

La obra reúne a especialistas de referencia en medicina, inteligencia artificial, derecho, bioética y universidad, ofreciendo una visión transversal, sólida y profundamente orientada al interés general. Sus capítulos analizan el presente y el futuro de la IA en medicina, su validación técnica y científica, su enseñanza en la profesión médica, el profesionalismo, la protección de datos, la responsabilidad y los derechos implicados en su desarrollo y aplicación.

Más que un diagnóstico del momento, este libro constituye una propuesta de marco para pensar y guiar el futuro. Lejos del entusiasmo superficial o de la alarma infundada, defiende una incorporación responsable de la inteligencia artificial, basada en la evidencia, la ética, la supervisión humana y la dignidad de la persona. Una publicación necesaria para una medicina que quiere innovar sin perder su fundamento humanista.

Libro Blanco de IA en Medicina

commálaga
Colegio de Médicos

commálaga
health *hub*