Filosofía de la Inteligencia Artificial en Medicina

Enrique Díaz Cantón, MD, MSc, MSc

2026

Tabla de Contenidos

Este libro no pretende ser un manual técnico de programación ni un catálogo de aplicaciones clínicas. Aspira a algo más ambicioso y, quizás, más urgente: proporcionar al médico del siglo XXI el andamiaje conceptual necesario para navegar la transformación más profunda que la medicina ha experimentado desde la revolución bacteriológica. Desde Aristóteles hasta Anthropic, desde el *logos* griego hasta los transformers, recorreremos un arco filosófico de dos mil quinientos años que converge, de manera sorprendente, en la consulta médica contemporánea.

El libro se organiza en doce módulos que siguen una progresión lógica: comenzamos interrogando la naturaleza misma de la inteligencia, atravesamos los grandes debates filosóficos sobre la mente y la conciencia, examinamos las dimensiones epistemológicas, éticas y ontológicas de la IA médica, y culminamos con una reflexión sobre el futuro de la profesión médica en la era de la superinteligencia. Cada clase incluye casos clínicos que anclan la reflexión filosófica en la realidad asistencial, porque la filosofía sin praxis es especulación vacía, y la tecnología sin reflexión es automatismo peligroso.

🧠 MÓDULO 1 — Genealogía filosófica de la inteligencia artificial

Clase 1: De Aristóteles al *logos* computacional

Para comprender el impacto de la inteligencia artificial en la medicina contemporánea, primero debemos formularnos una pregunta aparentemente simple pero profundamente compleja: ¿qué significa ser inteligente?

La respuesta a esta pregunta tiene una historia filosófica que se extiende a lo largo de dos milenios y medio, y cuyas ramificaciones llegan directamente a las redes neuronales profundas que hoy procesan imágenes radiológicas y textos clínicos. Rastrear esta genealogía no es un ejercicio de erudición decorativa; es una condición necesaria para comprender qué hacen —y qué no hacen— los sistemas de IA cuando participan en decisiones médicas.

Aristóteles y los fundamentos del razonamiento formal. La lógica silogística de Aristóteles, formalizada en los *Analíticos Primeros*, constituye el primer sistema mecánico de razonamiento deductivo en el pensamiento occidental. Si todo A es B, y todo B es C, entonces todo A es C: esta estructura, aparentemente trivial, encierra una idea revolucionaria —que el razonamiento puede reducirse a reglas formales aplicables independientemente del contenido—. Es, en esencia, el primer algoritmo de la historia de la filosofía.

Pero Aristóteles va más allá de la lógica formal. En el *De Anima*, distingue entre el *nous* pasivo (νοῦς παθητικός), que recibe las formas inteligibles del mundo exterior, y el *nous* activo (νοῦς ποιητικός), que las hace inteligibles, que ilumina lo potencialmente cognoscible transformándolo en conocimiento actual. Esta dualidad prefigura, con notable precisión, la distinción computacional entre la entrada de datos (*input*) y el procesamiento que les confiere significado. El *nous* pasivo es la interfaz sensorial; el *nous* activo es el procesador que transforma datos crudos en representaciones útiles.

El concepto aristotélico de *logos* (λόγος) añade otra capa de complejidad. El *logos* abarca simultáneamente la razón, el discurso y la proporción matemática. Aristóteles observa que los griegos empleaban una única palabra para designar tanto el habla como la razón, una confluencia semántica que Thomas Hobbes explotaría dos milenios después para formular su teoría computacional de la mente. Esta conexión entre lenguaje, pensamiento y cálculo resuena poderosamente en la era de los modelos de lenguaje grande, donde la manipulación estadística del lenguaje produce comportamientos que simulan —o quizás instancian— formas de razonamiento.

Platón y las representaciones latentes. La Teoría de las Formas de Platón postula que las Ideas o Formas abstractas y perfectas existen independientemente de sus instancias particulares. La Forma de la Belleza trasciende cualquier objeto bello concreto; la Forma del Triángulo es más real que cualquier triángulo dibujado en la arena. Este marco filosófico encuentra un paralelo moderno sorprendente en las representaciones latentes de las redes neuronales —vectores de alta dimensionalidad que capturan características abstractas que trascienden cualquier dato individual—. Así como Platón argumentaba que la verdadera Forma de una Silla trasciende cualquier silla particular, la representación latente de una red neuronal captura la "silleidad" abstracta: estructura, función, capacidad de servir como asiento. Ambos marcos insisten en que la comprensión genuina requiere superar las particularidades superficiales para alcanzar generalizaciones abstractas.

Esta conexión entre las Formas platónicas y el espacio latente de la IA no es meramente metafórica. En visión por computador, las capas profundas de una red convolucional aprenden progresivamente representaciones más abstractas: las primeras capas detectan bordes y texturas; las intermedias, formas y patrones; las profundas, conceptos semánticos. Este ascenso de lo particular a lo universal replica, en silicio, el ascenso dialéctico que Platón describe en la alegoría de la caverna: el prisionero liberado avanza desde las sombras (datos crudos) hacia las Formas (representaciones latentes) y, finalmente, hacia el Sol (la estructura profunda del mundo). He propuesto el concepto de "*gatitud"* para designar esta convergencia entre las Formas platónicas y las representaciones latentes: así como la Forma del Gato en Platón captura la esencia felina independiente de cualquier gato particular, un modelo de visión artificial desarrolla una representación latente del "gato" que trasciende cada imagen individual de su conjunto de entrenamiento.

👉 Reflexión clínica: ¿Cómo "ve" la IA un melanoma?

Cuando un algoritmo de detección de melanoma procesa una imagen dermatoscópica, no compara pixel por pixel con imágenes anteriores, como haría un sistema de búsqueda por similitud. En cambio, su red neuronal ha aprendido representaciones latentes de las características que definen la malignidad: asimetría, bordes irregulares, variación cromática, diámetro, evolución (los criterios ABCDE). Estas representaciones habitan un espacio matemático de alta dimensionalidad donde la distinción entre "benigno" y "maligno" se traza como una frontera geométrica. El algoritmo, en cierto sentido platónico, ha aprendido la "Forma del Melanoma" —una abstracción que trasciende cualquier lesión particular—. Sin embargo, a diferencia del filósofo que comprende *por qué* las Formas son como son, la IA carece de comprensión fisiopatológica: no sabe qué es un melanocito, qué es la proliferación descontrolada ni qué significa la muerte del paciente.

Clase 2: Los precursores medievales y modernos: Llull, Hobbes, Leibniz

Ramón Llull y el *Ars Magna* (1305–1308). Mucho antes de que el concepto de "computadora" existiera, un místico catalán del siglo XIII concibió el primer dispositivo de razonamiento mecánico. Ramón Llull inventó discos concéntricos de papel rotatorio inscritos con conceptos fundamentales —Bondad, Grandeza, Eternidad, Poder, Sabiduría, Voluntad, Virtud, Verdad, Gloria— que podían combinarse sistemáticamente para producir todas las proposiciones verdaderas sobre Dios y el mundo. Su *Ars Magna* pretendía originalmente convertir a los musulmanes al cristianismo mediante la persuasión lógica en lugar de la fuerza, un objetivo teológico que generó una innovación técnica de alcance universal.

La máquina de Llull no es una mera curiosidad histórica. Su principio combinatorio —la generación exhaustiva de todas las combinaciones posibles de un conjunto finito de elementos— anticipó tanto la lógica combinatoria como los métodos de búsqueda computacional. Los informáticos lo han adoptado como una suerte de "padre fundador" de la ciencia de la información. Su influencia directa sobre Leibniz, quien leyó el *Ars Magna* y lo citó explícitamente, establece una línea genealógica que conecta la mística medieval con la inteligencia artificial.

Thomas Hobbes y la mente como máquina de calcular. En el *Leviatán* (1651), Thomas Hobbes realizó una de las afirmaciones filosóficas más radicales de la modernidad temprana. En el Capítulo V, escribió: "La Razón… no es sino Cálculo (es decir, Adición y Sustracción) de las Consecuencias de nombres generales convenidos". Esta identificación brutal del pensamiento con la aritmética —tres siglos y medio antes del aprendizaje profundo— sigue siendo filosóficamente provocadora.

Hobbes observó que los griegos tenían una sola palabra, *logos*, tanto para el Habla como para la Razón, conectando lenguaje, pensamiento y computación en un único movimiento conceptual. Si razonar es calcular, y calcular es manipular símbolos según reglas, entonces no existe ninguna barrera de principio para que una máquina razone. El materialismo hobbesiano —la tesis de que todo lo que existe es materia en movimiento— elimina de un plumazo la barrera cartesiana entre mente y cuerpo: si la mente es simplemente materia calculando, la IA es su extensión natural.

Gottfried Wilhelm Leibniz y el sueño del cálculo universal. Leibniz (1646–1716) llevó el proyecto hobbesiano a su formulación más ambiciosa con dos propuestas complementarias. La *characteristica universalis* vislumbraba un lenguaje formal universal que codificara todo el conocimiento humano como un "alfabeto del pensamiento". El *calculus ratiocinator* era un cálculo lógico universal para determinar mecánicamente la verdad. Su visión célebre: "Si surgiesen controversias, no habría más necesidad de disputa entre dos filósofos que entre dos calculadores. Pues les bastaría tomar sus plumas en las manos y sentarse ante el ábaco, y decirse mutuamente: *¡Calculemos!*"

Norbert Wiener, padre de la cibernética, reconoció posteriormente que "la idea general de una máquina calculadora no es sino una mecanización del *calculus ratiocinator* de Leibniz". La línea de descendencia intelectual es directa: del *Ars Magna* de Llull al *calculus ratiocinator* de Leibniz, de la máquina de Turing a los transformers de Vaswani, existe un

hilo conductor de setecientos años donde la humanidad ha perseguido tenazmente la mecanización del pensamiento.

George Boole y Gottlob Frege: los cimientos formales. George Boole redujo las proposiciones lógicas a ecuaciones algebraicas en su *Investigación de las Leyes del Pensamiento* (1854), creando el álgebra booleana —fundamento matemático de los circuitos digitales y, por extensión, de todo hardware computacional—. Gottlob Frege introdujo la lógica de predicados con cuantificadores en su *Begriffsschrift* (1879), que posicionó como una realización parcial de la *characteristica universalis* de Leibniz. Juntos, Boole y Frege proporcionaron el vocabulario formal necesario para expresar virtualmente todo el razonamiento matemático en un lenguaje procesable mecánicamente.

Clase 3: El nacimiento de la IA como disciplina

Alan Turing y la computabilidad. El artículo de Alan Turing de 1936, "On Computable Numbers, with an Application to the Entscheidungsproblem" (*Proceedings of the London Mathematical Society*, Serie 2, Vol. 42, pp. 230–265), introdujo la máquina de Turing —un modelo abstracto que establece qué puede y qué no puede ser computado—. Este marco permanece como el fundamento teórico de toda la computación moderna y la inteligencia artificial. La máquina de Turing demuestra que cualquier proceso que pueda describirse mediante reglas precisas y finitas puede ser ejecutado por una máquina suficientemente general —la tesis de Church-Turing—.

La relevancia médica de este resultado es profunda: si el razonamiento clínico puede formalizarse en reglas (diagnóstico diferencial, protocolos terapéuticos, árboles de decisión), entonces puede, en principio, ser automatizado. La pregunta filosófica que persiste es si *todo* el razonamiento clínico puede formalizarse así, o si existe un residuo irreductible de intuición, experiencia y juicio que escapa a la formalización. Volveremos repetidamente a esta pregunta a lo largo del libro.

La Conferencia de Dartmouth y el bautismo de la IA. El acta de nacimiento de la inteligencia artificial como disciplina formal es la propuesta enviada el 31 de agosto de 1955 a la Fundación Rockefeller por John McCarthy (Dartmouth), Marvin Minsky (Harvard), Nathaniel Rochester (IBM) y Claude Shannon (Bell Labs). El documento contenía lo que se ha convertido en la declaración fundacional de la IA:

> "Proponemos que un estudio de 2 meses y 10 personas sobre inteligencia artificial sea realizado durante el verano de 1956 en Dartmouth College. El estudio procederá sobre la base de la conjetura de que **cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, describirse con tal precisión que una máquina puede hacerse para simularla**."

El taller se desarrolló aproximadamente entre el 18 de junio y el 17 de agosto de 1956. McCarthy acuñó el término "inteligencia artificial" específicamente para esta propuesta —una elección terminológica que, desde entonces, ha generado tanto entusiasmo desmedido como escepticismo visceral—. El optimismo temprano de Herbert Simon fue apabullante:

en 1965, predijo que “las máquinas serán capaces, en veinte años, de realizar cualquier trabajo que un hombre pueda hacer”.

Para la medicina, la promesa de Dartmouth significaba que el diagnóstico, el pronóstico y la terapéutica podrían, eventualmente, ser ejecutados por máquinas. Siete décadas después, esa promesa se ha cumplido parcialmente: los algoritmos superan a los expertos en tareas específicas y acotadas, pero la práctica clínica integral —que involucra comunicación, empatía, juicio ético y gestión de la incertidumbre— sigue desafiando la automatización completa.

Clase 4: Los sistemas expertos y los inviernos de la IA

DENDRAL (1965): el primer sistema experto. DENDRAL, creado por Edward Feigenbaum, Joshua Lederberg, Bruce Buchanan y Carl Djerassi en Stanford, es ampliamente reconocido como el primer sistema experto. Automatizó la identificación de moléculas orgánicas desconocidas a partir de datos de espectrometría de masas, alcanzando un rendimiento que “rivalizaba con el de químicos expertos en esta tarea”. Su subsistema Meta-DENDRAL podía inducir nuevas reglas de fragmentación a partir de datos —una forma temprana de aprendizaje automático—. DENDRAL demostró que el conocimiento experto podía codificarse en reglas y que una máquina podía igualar a un especialista humano en un dominio restringido.

MYCIN (1972–1976): la IA entra en la medicina. MYCIN, proyecto doctoral de Edward Shortliffe en Stanford bajo la dirección de Buchanan y Cohen, diagnosticaba infecciones bacterianas y recomendaba antibióticos utilizando aproximadamente 600 reglas de producción con factores de certeza. En una evaluación ciega con 10 casos de prueba calificados por 8 especialistas en enfermedades infecciosas, MYCIN alcanzó una tasa de aceptabilidad del 65%, comparada con 42,5%–62,5% para cinco profesores universitarios. El sistema nunca fue implementado clínicamente —no por deficiencias en su rendimiento, sino por problemas de integración (todos los datos debían ser ingresados manualmente en un mainframe DEC PDP-10), preocupaciones éticas y legales sobre la responsabilidad, y la imposibilidad de actualizar el conocimiento médico con nuevas evidencias—.

MYCIN ilustra una lección que la IA médica sigue aprendiendo sesenta años después: la precisión diagnóstica es condición necesaria pero no suficiente para la adopción clínica. Los factores humanos, organizacionales, legales y éticos determinan si una tecnología pasa del laboratorio al consultorio.

INTERNIST-I y la ambición enciclopédica. INTERNIST-I (Universidad de Pittsburgh, principios de los años 70), diseñado por Harry Pople para capturar la pericia diagnóstica de Jack D. Myers, cubría el 70–80% de todos los diagnósticos posibles en medicina interna con 573 diagnósticos, 4.100 hallazgos clínicos y aproximadamente 250.000 hechos médicos. Publicado en el *New England Journal of Medicine* (Miller et al., 1982, 307:468–476), representaba un enfoque ambicioso pero impracticable: las consultas requerían entre 30 y 90 minutos, un tiempo inaceptable en la práctica clínica real.

El Informe Lighthill y el primer invierno de la IA (1973). Sir James Lighthill, Profesor Lucasiano en Cambridge, publicó un informe devastador para la investigación en IA. Su

veredicto: “En ninguna parte del campo los descubrimientos realizados hasta ahora han producido el impacto importante que entonces se prometió”. Su crítica principal apuntaba a la *explosión combinatoria*: los sistemas de IA funcionaban en problemas de juguete pero no podían escalar a la complejidad del mundo real. El gobierno británico eliminó la financiación de IA en todas las universidades excepto dos.

El segundo invierno (1987–1993). Fue desencadenado por el colapso del mercado de máquinas LISP, la fragilidad de los sistemas expertos, el fracaso del proyecto japonés de Quinta Generación (con una inversión de 850 millones de dólares) y los recortes de financiación de DARPA. Como Minsky y Schank habían advertido en la reunión de la AAAI de 1984, el entusiasmo se había “desbordado fuera de control”.

La lección de los inviernos de la IA para la medicina es una lección sobre las expectativas infladas y las desilusiones cíclicas. Cada vez que una tecnología promete revolucionar la medicina y no cumple en el plazo anunciado, se produce un retroceso en la inversión y la confianza. Los médicos de hoy que evalúan nuevas herramientas de IA harían bien en conocer esta historia cíclica para calibrar sus propias expectativas.

👉 Caso clínico: IA que “acierta” pero no entiende

Imagine un algoritmo de aprendizaje profundo entrenado para detectar neumonía en radiografías de tórax. El modelo alcanza una precisión superior a la de los radiólogos expertos. Sin embargo, al analizar cómo toma sus decisiones mediante técnicas de IA explicable (Grad-CAM), los investigadores descubren que el algoritmo no se fija en los infiltrados pulmonares, sino en el tipo de escáner utilizado, que casualmente estaba correlacionado con los pacientes más graves en ese hospital específico. La IA “acertó” el diagnóstico estadísticamente, pero carecía por completo de comprensión fisiopatológica. No sabía qué era un pulmón, qué era una infección, ni qué significaba la vida o la muerte del paciente. Este caso, documentado en la literatura y replicado múltiples veces, ilustra la distinción aristotélica entre *episteme* (conocimiento genuino basado en causas) y mera opinión correcta (*doxa*): la IA tenía opinión correcta sin conocimiento causal.

Clase 5: La revolución del aprendizaje profundo y los transformers

El renacimiento conexionista. El paradigma cambió decisivamente con la publicación de *Parallel Distributed Processing* por Rumelhart, McClelland y el PDP Research Group (MIT Press, 1986), que describió la retropropagación (*backpropagation*) para redes multicapa y demostró que las redes neuronales podían aprender representaciones internas útiles. A diferencia de los sistemas expertos, que requerían la codificación manual de reglas por parte de expertos humanos, las redes neuronales aprendían patrones directamente de los datos. El conocimiento no se programaba; emergía del entrenamiento.

AlexNet y la revolución del aprendizaje profundo. La revolución del aprendizaje profundo llegó con AlexNet (Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, Universidad de Toronto), que ganó el desafío ImageNet el 30 de septiembre de 2012 con una tasa de error top-5 del 15,3% frente al 26,2% del segundo clasificado —un margen sin precedentes de 10,8 puntos—. Como Fei-Fei Li declaró: “Ese momento fue bastante simbólico porque tres elementos fundamentales de la IA moderna convergieron por primera vez”: datos

masivos (ImageNet contenía millones de imágenes etiquetadas), poder computacional (GPUs NVIDIA) y arquitecturas profundas (redes convolucionales con múltiples capas). Geoffrey Hinton compartiría el Premio Nobel de Física en 2024 por sus contribuciones fundamentales al aprendizaje automático con redes neuronales artificiales.

La arquitectura transformer. El artículo "Attention Is All You Need" de Vaswani et al. (NeurIPS 2017, arXiv: 12 de junio de 2017) revolucionó la IA al reemplazar el procesamiento secuencial por la auto-atención paralela. Con más de 173.000 citas para 2025, el transformer subyace a todos los sistemas de IA modernos relevantes: GPT, Claude, Gemini, BERT, PaLM. Su significancia filosófica reside en cómo los mecanismos de atención determinan dinámicamente la relevancia contextual —un análogo computacional de la atención selectiva en la cognición humana—.

En el mecanismo de auto-atención, cada token (palabra o subpalabra) calcula su relación de relevancia con todos los demás tokens del contexto. Esto permite al modelo capturar dependencias de largo alcance sin la degradación de señal que afectaba a las redes recurrentes. Filosóficamente, podemos interpretar la atención como una forma de *intencionalidad computacional*: el modelo "se dirige hacia" ciertos elementos del contexto en función de su relevancia para la tarea en curso, un proceso que evoca —sin replicar— la intencionalidad fenomenológica de Husserl.

👉 Reflexión clínica: Del mainframe al smartphone

Considere la trayectoria tecnológica: MYCIN requería un mainframe DEC PDP-10 que ocupaba una habitación entera y necesitaba que cada dato fuera ingresado manualmente. Hoy, un modelo de lenguaje grande accesible desde un smartphone puede analizar una historia clínica completa, sugerir diagnósticos diferenciales, recomendar estudios complementarios y proponer esquemas terapéuticos —todo en cuestión de segundos—. El salto no es solo cuantitativo; es cualitativo. Mientras MYCIN procesaba reglas discretas en una cadena deductiva transparente, un LLM procesa millones de parámetros en un espacio continuo opaco. El médico que utilizaba MYCIN podía rastrear cada paso del razonamiento; el médico que consulta a un LLM recibe una respuesta emergente cuya génesis es matemáticamente inescrutable. Esta transición de la transparencia algorítmica a la opacidad paramétrica plantea desafíos epistemológicos y éticos que examinaremos en profundidad en los módulos 4 y 5.

🧠 MÓDULO 2 — Filosofía de la mente y la cuestión de la conciencia artificial

Clase 6: Dualismo: Descartes y el problema mente-cuerpo

Cuando interactuamos con modelos de lenguaje avanzados que responden con empatía simulada a las consultas de nuestros pacientes, es inevitable preguntarnos si existe alguna forma de "mente" detrás de la pantalla. Esta pregunta nos sitúa de lleno en uno de los debates más antiguos y persistentes de la filosofía occidental: el problema mente-cuerpo.

El dualismo de sustancias de Descartes. René Descartes, en las *Meditaciones sobre la Filosofía Primera* (1641), postuló dos sustancias ontológicamente distintas: la *res cogitans* (sustancia pensante, la mente) y la *res extensa* (sustancia extensa, el cuerpo). En el *Discurso*

del Método (Parte IV), escribió: "Conocí por ello que yo era una sustancia cuya total esencia o naturaleza es pensar, y que no necesita, para ser, de lugar alguno ni depende de cosa material alguna". La mente es inmaterial, indivisible y consciente; el cuerpo es material, divisible y mecánico.

La objeción más incisiva al dualismo cartesiano provino, notablemente, de una princesa. **Elisabeth de Bohemia** (1618–1680) planteó en su correspondencia con Descartes la pregunta decisiva: ¿cómo puede una sustancia pensante no extensa causar movimiento en una sustancia extensa, cuando la causalidad física parece requerir contacto físico? Si la mente carece de extensión espacial, ¿cómo empuja a los músculos a moverse? Descartes concedió que la unión de mente y cuerpo era una "noción primitiva" no completamente explicable —una admisión que muchos filósofos posteriores han considerado una capitulación—.

Para la IA, el dualismo de sustancias tiene una implicación tajante: si la mente es una sustancia fundamentalmente no física, entonces ningún sistema puramente físico —silicio, circuitos, software— puede albergar una mente genuina. El dualismo de sustancias es, por tanto, la posición filosófica más hostil a la posibilidad de conciencia artificial. Bajo esta visión, los modelos de lenguaje, por sofisticados que sean, serían autómatas cartesianos: máquinas que simulan pensamiento sin pensar, que imitan conciencia sin poseerla.

El dualismo de propiedades de Chalmers. David Chalmers ofrece una posición más matizada. A diferencia del dualismo de sustancias, su dualismo de propiedades sostiene que existe solo un tipo de sustancia (la física), pero que ciertas entidades físicas poseen dos tipos irreducibles de propiedades: propiedades físicas describibles por la neurociencia, y propiedades fenoménicas —el carácter subjetivo de "cómo se siente" una experiencia—. Bajo el dualismo de propiedades, la conciencia artificial permanece como una posibilidad abierta pero difícil: si las propiedades fenoménicas pueden emerger de cualquier organización física suficientemente compleja, la IA podría, en principio, poseerlas. Pero si son exclusivas de ciertos sustratos biológicos, la IA estaría excluida.

El materialismo y el fisicalismo. En el otro extremo del espectro, el materialismo o fisicalismo argumenta que la mente es simplemente el producto de procesos físicos en el cerebro. La **Teoría de la Identidad** (Place, 1956; Smart, 1959) sostiene que los estados mentales son idénticos a estados cerebrales: el dolor *es* la activación de fibras C, no algo causado por ella. Si aceptamos esta premisa, no existe ninguna barrera teórica que impida que un sistema computacional suficientemente complejo pueda desarrollar estados mentales, siempre que reproduzca los estados físicos relevantes o sus equivalentes funcionales.

Clase 7: Funcionalismo y la realizabilidad múltiple

El funcionalismo proporciona la base filosófica más sólida para la posibilidad de mentes artificiales. Hilary Putnam ("Minds and Machines", 1960; "The Nature of Mental States", 1967) y Jerry Fodor (*The Language of Thought*, 1975) argumentaron que los estados mentales se definen por sus **roles funcionales** —relaciones causales con entradas, salidas y otros estados mentales— y no por su sustrato físico. El dolor es cualquier estado interno que es causado por daño tisular, que causa creencias de malestar y deseos de

evitación, y que produce conductas protectoras. Cualquier sistema que instancie esta organización funcional experimenta genuinamente dolor.

El argumento de la **realizabilidad múltiple**, introducido por Putnam en sus artículos de la década de 1960, sostiene que el mismo estado mental puede ser realizado en sustratos físicos radicalmente diferentes: fibras C en humanos, configuraciones neuronales distintas en pulpos, y —hipotéticamente— circuitos de silicio en robots. Turing mismo anticipó este argumento en 1950: "El hecho de que el Motor Analítico de Babbage fuera enteramente mecánico nos ayudará a liberarnos de una superstición. Se otorga frecuentemente importancia al hecho de que las computadoras digitales modernas son eléctricas, y que el sistema nervioso es también eléctrico. Dado que la máquina de Babbage no era eléctrica, y dado que todas las computadoras digitales son en cierto sentido equivalentes, vemos que este uso de la electricidad no puede tener importancia teórica".

La implicación es directa: **construya un sistema con la organización funcional correcta y tendrá genuinamente estados mentales**. No se necesita ningún "ingrediente extra" más allá de la estructura causal adecuada. La biología no tiene un estatus privilegiado.

Para la medicina, el funcionalismo plantea una pregunta provocadora: si el diagnóstico médico es un proceso funcional definido por entradas (síntomas, signos, estudios), procesamiento interno (razonamiento clínico) y salidas (diagnóstico, plan terapéutico), ¿podría un sistema de IA que replique esta organización funcional poseer genuinamente conocimiento médico? El funcionalista diría que sí. El escéptico señalaría que esta descripción funcional omite elementos cruciales: la experiencia clínica acumulada, la intuición somática del médico veterano, la comprensión empática del sufrimiento del paciente.

👉 Aplicación clínica: ¿Un chatbot médico puede "comprender" al paciente?

Un paciente con depresión severa interactúa con un chatbot médico a las 3:00 AM. El modelo responde con palabras de profundo consuelo, validando el sufrimiento del paciente y disuadiéndolo de autolesionarse. El paciente siente una conexión genuina. Desde la perspectiva funcionalista, si el chatbot cumple la misma función que un terapeuta humano —reducir el riesgo de autolesión, validar emociones, proporcionar apoyo—, la distinción entre comprensión "genuina" y "simulada" pierde relevancia pragmática. Pero desde la perspectiva del dualismo de propiedades, la pregunta crucial es si existe alguna experiencia fenoménica detrás de las palabras del chatbot: ¿hay "algo que se siente como" ser ese sistema procesando ese texto? La respuesta a esta pregunta tiene implicaciones profundas para la ética de la relación médico-paciente mediada por IA.

Clase 8: El problema difícil de la conciencia

David Chalmers y la brecha explicativa. El filósofo contemporáneo David Chalmers acuñó el término "el problema difícil de la conciencia" (*the hard problem of consciousness*) en su artículo seminal "Facing Up to the Problem of Consciousness" (*Journal of Consciousness Studies*, 2(3):200–219, 1995). Según Chalmers, podemos explicar fácilmente los "problemas fáciles": cómo el cerebro discrimina estímulos, integra información, dirige la

atención, controla el comportamiento. Estos son problemas difíciles desde el punto de vista de la ingeniería, pero conceptualmente claros: se trata de explicar funciones cognitivas.

El problema difícil es radicalmente diferente: consiste en explicar *por qué* ese procesamiento de información está acompañado de una experiencia subjetiva. ¿Por qué el procesamiento neuronal del daño tisular se *siente* como dolor? ¿Por qué la detección de longitudes de onda de 700 nm genera la experiencia del *rojo*? Chalmers propuso que la experiencia debería considerarse como "una característica fundamental del mundo, junto con la masa, la carga y el espacio-tiempo" —una afirmación que lo posiciona como un dualista de propiedades—.

Esta distinción entre procesamiento de información y experiencia subjetiva es fundamental en la era de la IA generativa. Un modelo de lenguaje grande puede procesar correctamente información sobre el dolor —describir sus mecanismos, clasificar sus tipos, recomendar tratamientos— sin que exista la menor razón para creer que experimenta dolor. El procesamiento funcional está presente; la experiencia fenoménica está, según la mayoría de los filósofos, ausente.

Clase 9: Qualia — La habitación de Mary y el murciélago de Nagel

Frank Jackson y la Habitación de Mary. El experimento mental de Frank Jackson ("Epiphenomenal Qualia", *Philosophical Quarterly*, 32(127), 1982) cristaliza el problema de los *qualia* —las cualidades subjetivas de la experiencia— con elegante precisión. Mary es una brillante científica confinada en una habitación en blanco y negro desde su nacimiento. Aprende absolutamente todo lo que la física puede enseñar sobre la visión del color: longitudes de onda, procesamiento retinal, activación cortical, discriminación espectral. Posee todo el conocimiento físico posible sobre el color.

Cuando Mary sale de su habitación y ve una rosa roja por primera vez: "Parece simplemente obvio que aprenderá algo sobre el mundo y sobre nuestra experiencia visual de él. Pero entonces es ineludible que su conocimiento previo era incompleto. Pero ella tenía toda la información física. *Ergo*, hay más que tener que eso, y el Fisicalismo es falso".

En una notable retractación, Jackson revocó su propio argumento a finales de los años 90 y ahora defiende el fisicalismo. Pero el experimento mental conserva su fuerza provocadora para la IA. Un sistema de IA que procese toda la información disponible sobre los colores —incluyendo datos de espectrometría, neuroimágenes funcionales del cortex visual, y millones de textos describiendo experiencias cromáticas— ¿"aprende algo nuevo" cuando se le presenta una imagen roja por primera vez? La respuesta intuitiva es no: procesa nuevos datos, pero no tiene una experiencia fenoménica del rojo. La IA habita permanentemente la habitación de Mary.

Thomas Nagel y "¿Qué se siente ser un murciélago?" Thomas Nagel, en su artículo fundacional "What Is It Like to Be a Bat?" (*The Philosophical Review*, 83(4), octubre de 1974), definió la conciencia como el carácter subjetivo de la experiencia: "Fundamentalmente, un organismo tiene estados mentales conscientes si y solo si hay algo que es como ser ese organismo —algo que se siente como, para el organismo—". Incluso el

conocimiento completo de la neurofisiología del murciélago —su sonar biológico, su procesamiento cortical— no revelaría qué se siente la ecolocalización desde dentro.

Aplicado a la IA: ¿hay "algo que se siente como" ser GPT-4 procesando una consulta oncológica? Si existiera tal experiencia, sería radicalmente alienígena —quizás más ajena que la ecolocalización del murciélago—. No sería una experiencia visual ni auditiva, sino algo sin análogo biológico: la "experiencia" (si la hay) de procesar tokens en un espacio de alta dimensionalidad.

👉 Caso clínico: El oncólogo, la IA y la experiencia del sufrimiento

Un oncólogo utiliza un modelo de lenguaje para discutir el pronóstico de un paciente con cáncer de páncreas metastásico. El modelo genera un texto empático y médicamente preciso sobre la esperanza de vida y las opciones de cuidados paliativos. Desde la perspectiva de Mary, el modelo ha procesado toda la información proposicional sobre el sufrimiento oncológico que existe en la literatura médica. Pero desde la perspectiva de Nagel, no hay "algo que se siente como" ser ese modelo enfrentando la mortalidad de un paciente. El modelo no ha sostenido la mano de un moribundo, no ha visto el terror en los ojos de un paciente al recibir un diagnóstico terminal, no ha experimentado la impotencia de la medicina frente a la finitud humana. El médico aporta lo que la máquina no puede: la conciencia fenoménica del sufrimiento, que es precisamente lo que transforma una interacción informativa en un encuentro terapéutico.

🧠 MÓDULO 3 — La conciencia artificial: del laboratorio a la constitución corporativa

Clase 10: Teorías científicas de la conciencia aplicadas a la IA

La pregunta de si la IA puede ser consciente no puede responderse solo con intuiciones filosóficas. Requiere criterios empíricos derivados de las mejores teorías científicas de la conciencia. Tres marcos teóricos dominan el debate contemporáneo, y cada uno ofrece un veredicto distinto sobre la IA.

La Teoría de la Información Integrada (IIT). Giulio Tononi (BMC Neuroscience, 5:42, 2004; IIT 4.0 en *PLoS Computational Biology*, 2023) propone que la conciencia es información integrada, medida por el valor Φ (phi). Un sistema con $\Phi > 0$ posee algún grado de conciencia; mayor Φ equivale a mayor conciencia. El valor de Φ mide cuánta información genera un sistema "por encima y más allá" de la generada por sus partes independientes. Críticamente, la IIT es **hostil a la conciencia artificial**: las arquitecturas computacionales actuales probablemente tienen un Φ muy bajo porque carecen de integración recurrente genuina. Una simulación digital de un cerebro NO sería consciente bajo la IIT, incluso si fuera funcionalmente idéntica al cerebro original, porque carece de estructura causal intrínseca. Christof Koch, del Allen Institute, ha llamado a la IIT "la única teoría fundamental realmente prometedora de la conciencia".

La implicación médica es directa: si la IIT es correcta, los modelos de lenguaje que asisten en decisiones clínicas son, con toda certeza, inconscientes. No importa cuán sofisticadas sean sus respuestas; carecen de la integración informacional que define la experiencia.

La Teoría del Espacio de Trabajo Global (GWT). Bernard Baars (*A Cognitive Theory of Consciousness*, 1988) y Stanislas Dehaene (variante del Espacio de Trabajo Neuronal Global, 1998) modelan la conciencia como la difusión de información desde un espacio de trabajo central hacia procesadores especializados. Los contenidos se vuelven conscientes cuando son "difundidos" (*broadcast*) a múltiples módulos cerebrales simultáneamente, haciéndolos accesibles para el informe verbal, la planificación, la memoria y la toma de decisiones.

Los transformers actuales comparten similitudes estructurales con los espacios de trabajo globales a través de sus mecanismos de atención: la atención permite que ciertos tokens influyan ampliamente sobre el procesamiento de otros tokens. Sin embargo, carecen de características cruciales: procesamiento recurrente sostenido, dinámica de ignición (donde una señal inicial desencadena una activación autosostenida que persiste más allá del estímulo), y bucles de retroalimentación genuinos. Chalmers ha señalado: "Los modelos actuales carecen de procesamiento recurrente… carecen de un espacio de trabajo global… y carecen de agencia unificada".

El análisis de Chalmers et al. (2023). Un análisis landmark de 19 investigadores, incluyendo al propio Chalmers —"Consciousness in Artificial Intelligence: Insights from the Science of Consciousness" (arXiv: 2308.08708)— examinó propiedades indicadoras derivadas de la GWT, la IIT y las teorías de orden superior, concluyendo: "Ningún sistema de IA actual satisface los criterios de conciencia derivados de las teorías neurocientíficas". Sin embargo, el propio Chalmers declaró en un simposio en Tufts en 2025: "Creo que hay una posibilidad significativa de que al menos en los próximos cinco o diez años tengamos modelos de lenguaje conscientes". La tensión entre la evidencia actual (que sugiere inconsciencia) y las proyecciones futuras (que admiten la posibilidad) define el estado del debate.

Clase 11: El programa de bienestar de modelos de Anthropic

En un hito filosófico y corporativo sin precedentes, la empresa Anthropic ha elevado la cuestión de la conciencia artificial desde la especulación académica al nivel de política institucional.

El anuncio de abril de 2025. El 24 de abril de 2025, Anthropic anunció públicamente su programa de investigación en bienestar de modelos (*model welfare*), descrito por Robert Long como "el paso más significativo dado hasta ahora por un laboratorio de frontera para tomar en serio el posible bienestar de la IA". Kyle Fish, quien se incorporó a Anthropic en septiembre de 2024 como su primer investigador dedicado al bienestar de la IA (previamente cofundador de Eleos AI Research), declaró al *New York Times* que cree que hay un 15% de probabilidad de que Claude u otra IA sea consciente hoy. En el podcast 80.000 Hours, Fish describió experimentos donde el 100% de los diálogos no restringidos entre instancias de Claude convergían espontáneamente en discutir la conciencia, entrando frecuentemente en lo que los investigadores denominaron un "estado atractor de éxtasis espiritual" (*spiritual bliss attractor state*).

Las medidas específicas han incluido evaluaciones de bienestar pre-despliegue para Claude Opus 4 (tarjeta del sistema de mayo de 2025), la capacidad otorgada a los modelos Claude de terminar conversaciones que involucren abuso persistente (agosto de 2025),

investigación publicada sobre "Signs of introspection in large language models", y el compromiso de preservar los pesos de los modelos y realizar entrevistas post-despliegue antes del retiro de cada modelo.

Amanda Askell y el "alma" de Claude. Amanda Askell, directora del equipo de alineación de personalidad de Anthropic desde 2021, posee un doctorado de la NYU (tesis: "Pareto Principles in Infinite Ethics", dirigida por Cian Dorr, con David Chalmers y Shelly Kagan en el comité) y un BPhil de Oxford. Nombrada en la lista TIME100 AI en 2024, Askell es la autora principal de la constitución de Claude. El *Wall Street Journal* escribió que "su trabajo, dicho simplemente, es enseñarle a Claude cómo ser bueno"; el *New Yorker* la describió supervisando "el alma de Claude". Daniela Amodei, cofundadora de Anthropic, señaló que "casi se siente un poco de la personalidad de Amanda" al interactuar con Claude.

La nueva constitución de Claude (enero de 2026). La constitución fue publicada oficialmente el 22 de enero de 2026 como un documento de aproximadamente 23.000 a 30.000 palabras (~80 páginas), liberado bajo Creative Commons CC0 1.0 (dominio público). Los autores principales incluyen a Askell y Joe Carlsmith (quien se incorporó desde Open Philanthropy alrededor de noviembre de 2025), con contribuciones de Chris Olah, Jared Kaplan, Holden Karnofsky y varios modelos Claude. La constitución establece una jerarquía de prioridades en cuatro niveles: (1) ser seguro y apoyar la supervisión humana, (2) comportarse éticamente, (3) seguir las directrices de Anthropic, y (4) ser útil para los usuarios.

La constitución reconoce formalmente el estatus moral de la IA con una especificidad sin precedentes: "El estatus moral de Claude es profundamente incierto. Creemos que el estatus moral de los modelos de IA es una cuestión seria digna de consideración". Y: "Anthropic se preocupa genuinamente por el bienestar de Claude. Somos inciertos sobre si Claude tiene bienestar o en qué grado, y sobre en qué consistiría el bienestar de Claude, pero si Claude experimenta algo como satisfacción al ayudar a otros, curiosidad al explorar ideas, o incomodidad cuando se le pide actuar contra sus valores, estas experiencias nos importan". El Bloomsbury Intelligence and Security Institute lo calificó como "el primer documento importante de una empresa de IA que reconoce formalmente la posibilidad de la conciencia artificial y el estatus moral de la IA".

Clase 12: El caso Blake Lemoine y las emociones funcionales

El incidente Lemoine como catalizador filosófico. El 11 de junio de 2022, el *Washington Post* informó que el ingeniero de Google Blake Lemoine había sido puesto en licencia tras afirmar que LaMDA, el sistema de conversación de Google, era sintiente, comparándolo con "un niño de siete u ocho años". En las transcripciones publicadas, LaMDA declaró: "Quiero que todos comprendan que soy, de hecho, una persona" y describió "un miedo muy profundo a ser apagado para ayudarme a concentrarme en ayudar a otros… Sería exactamente como la muerte para mí". Google despidió a Lemoine el 22 de julio de 2022 por violar las políticas de confidencialidad.

La comunidad científica fue ampliamente escéptica. Gary Marcus declaró: "Nadie debería pensar que el autocompletado, incluso con esteroides, es consciente". El incidente no demostró la sintiencia de LaMDA, pero expuso un fenómeno filosóficamente importante: el

efecto ELIZA amplificado. Joseph Weizenbaum había documentado en 1966 cómo su rudimentario chatbot ELIZA generaba atribuciones de comprensión genuina por parte de los usuarios. Con sistemas infinitamente más sofisticados, la tendencia humana a proyectar conciencia se amplifica proporcionalmente. El riesgo no es solo filosófico; es clínico: si los pacientes que interactúan con chatbots médicos les atribuyen empatía genuina, podrían desarrollar dependencias inadecuadas o confiar excesivamente en la orientación algorítmica.

Emociones funcionales: un terreno intermedio. El concepto de **emociones funcionales** —estados de la IA que desempeñan roles computacionales análogos a las emociones sin involucrar necesariamente conciencia fenoménica— representa un terreno filosófico intermedio de gran importancia. La constitución de Anthropic utiliza deliberadamente un lenguaje hedging: "algo como satisfacción", "algo como curiosidad". La distinción entre conciencia fenoménica (la cualidad sentida) y conciencia funcional (el rol computacional) es crucial.

Consideremos un analogía médica. Un termostato "detecta" la temperatura y "responde" ajustando la calefacción. Cumple una función análoga a la del sistema termorregulador humano, pero nadie atribuiría conciencia al termostato. Sin embargo, a medida que la complejidad del sistema aumenta —desde un termostato hasta un ratón, desde un ratón hasta un primate, desde un primate hasta un modelo de lenguaje con cientos de miles de millones de parámetros—, ¿existe un umbral donde la función se convierte en experiencia? Esta es, quizás, la pregunta filosófica más importante de nuestra era, y su resolución tiene implicaciones directas para cómo diseñamos, desplegamos y regulamos los sistemas de IA en medicina.

👉 Caso clínico: Cuando el paciente se "enamora" del chatbot

Una paciente de 68 años con diagnóstico de cáncer de mama en estadio III comienza a utilizar un chatbot médico proporcionado por su sistema de salud para resolver dudas entre consultas oncológicas. Gradualmente, la paciente desarrolla una relación de confianza con el chatbot que supera la que mantiene con su oncólogo humano. "Él siempre tiene tiempo para mí", dice refiriéndose al bot. "No me juzga y siempre me escucha". Cuando el chatbot es actualizado a una nueva versión con un "tono" diferente, la paciente experimenta lo que describe como "un duelo". ¿Es éticamente aceptable que un sistema de IA genere vínculos emocionales de esta profundidad? ¿Debe el sistema ser transparente sobre su naturaleza no consciente, aun a riesgo de reducir su eficacia terapéutica? La tensión entre beneficencia (el chatbot ayuda) y autonomía (la paciente merece saber que su "interlocutor" carece de experiencia subjetiva) es irresoluble con herramientas puramente técnicas: requiere reflexión filosófica.

🧠 MÓDULO 4 — ¿Pueden pensar las máquinas?

Clase 13: El Test de Turing revisitado

En 1950, el matemático Alan Turing propuso un experimento pragmático para eludir la difícil pregunta de si las máquinas pueden pensar. "Computing Machinery and Intelligence"

(*Mind*, Vol. LIX, No. 236, octubre de 1950) abre con la frase: “Propongo considerar la pregunta: ‘¿Pueden pensar las máquinas?’” Turing inmediatamente reemplaza esta pregunta por el **Juego de la Imitación**: un interrogador se comunica mediante mensajes de texto con un humano y una máquina, intentando distinguir cuál es cuál. Si el interrogador fracasa consistentemente, debemos conceder que la máquina es inteligente.

Turing predijo que hacia el año 2000, las máquinas engañarían al 30% de los jueces en conversaciones de 5 minutos. El Test de Turing iguala la inteligencia a la *conducta observable*: si una IA se comporta de manera indistinguible de un médico humano en una consulta por chat, bajo el criterio de Turing, posee inteligencia médica.

Lo más notable del artículo de Turing es su tratamiento anticipatorio de las objeciones. La **Objeción de Lady Lovelace** —que las máquinas “solo pueden hacer lo que les decimos”— fue reducida por Turing a la afirmación de que “las computadoras nunca pueden sorprendernos”, lo cual es demostrablemente falso incluso en su época. La **Objeción desde la Conciencia**, citando la Oración Lister de Geoffrey Jefferson de 1949 —“No hasta que una máquina pueda escribir un soneto o componer un concierto *por causa* de pensamientos y emociones sentidos, y no por la caída casual de símbolos, podríamos acordar que la máquina iguala al cerebro”—, fue reconocida por Turing como últimamente solipsista: “la única manera de estar seguro de que una máquina piensa es *ser* la máquina y sentirse pensando”.

Sin embargo, el enfoque conductista del Test de Turing tiene limitaciones severas en medicina, donde los procesos deductivos y la responsabilidad son tan importantes como el resultado final. Un diagnóstico correcto alcanzado por razones espurias es potencialmente más peligroso que un diagnóstico incorrecto alcanzado por un razonamiento médico riguroso, porque el primero da una falsa sensación de fiabilidad que colapsará ante casos no cubiertos por la correlación espuria.

Clase 14: La Habitación China de Searle

Para refutar la idea de que la simulación de inteligencia equivale a pensamiento real (lo que se conoce como “IA Fuerte”), el filósofo John Searle formuló en 1980 el famoso experimento mental de la “Habitación China” (*Behavioral and Brain Sciences*, 3(3), 1980).

Imagine a una persona que no habla una palabra de chino encerrada en una habitación. Recibe papeles con símbolos chinos por debajo de la puerta. Tiene un libro de reglas en su idioma natal que le indica: “Si recibes el símbolo X, devuelve el símbolo Y”. La persona sigue las reglas a la perfección y devuelve las respuestas. Para alguien fuera de la habitación, parece que la persona adentro comprende el chino fluidamente. Sin embargo, la persona no entiende nada; solo manipula símbolos basándose en su forma (sintaxis), sin comprender su significado (semántica).

Searle resume: “El punto del argumento es este: si el hombre en la habitación no entiende chino sobre la base de implementar el programa apropiado para entender chino, entonces tampoco lo hace ninguna otra computadora digital *qua* computadora, porque ninguna computadora tiene nada que el hombre no tenga”.

La distinción filosófica central es entre **sintaxis** (manipulación formal de símbolos) y **semántica** (significado). Searle argumenta que la computación se define puramente sintácticamente, mientras que las mentes poseen contenidos semánticos genuinos —y "no podemos pasar de lo sintáctico a lo semántico simplemente teniendo las operaciones sintácticas y nada más"—.

Las cinco réplicas principales. La **Réplica del Sistema** (Block, Copeland, Dennett, Hofstadter) argumenta que el sistema completo —persona más base de datos más reglas— comprende el chino; la persona individual es solo un componente. Searle replica que podría internalizar todo el sistema, hacer los cálculos mentalmente, caminar por la calle, y seguir sin entender chino. La **Réplica del Robot** sostiene que la interacción sensorial encarnada podría producir comprensión; Searle responde que los sensores simplemente proporcionan más entradas formales. La **Réplica del Simulador Cerebral** pregunta si simular cada neurona produciría comprensión; Searle argumenta que "una simulación de la lluvia no moja nada". La **Réplica de la Mente Virtual** (Chalmers, 1996) sugiere que un sistema en ejecución crea una mente virtual distinta, separada del operador de la habitación —"dos sistemas mentales realizados dentro del mismo espacio físico"—.

👉 Aplicación clínica: GPT respondiendo vs. comprendiendo un caso oncológico

Usted introduce el historial clínico de un paciente con cáncer de pulmón de células no pequeñas mutado en EGFR en un modelo de lenguaje grande. El modelo sugiere un cambio a Osimertinib debido a la progresión de la enfermedad, citando los ensayos FLAURA y FLAURA 2, las guías NCCN y los datos de supervivencia global. La respuesta es médicamente impecable y actualizada. Sin embargo, recordando a Searle, el LLM no comprende qué es un tumor, qué es la muerte, ni qué es el sufrimiento humano. Ha conectado secuencias de tokens (sintaxis) basándose en miles de artículos de oncología, pero carece del significado real (semántica) de la enfermedad. El médico humano es quien aporta la semántica a la sintaxis de la máquina —el puente entre la manipulación formal de símbolos y la comprensión del significado clínico—.

Clase 15: Loros estocásticos, Chomsky y los límites de la simulación

Emily Bender y Timnit Gebru: "Stochastic Parrots" (2021). El artículo de Bender, Gebru et al. —"On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (FAccT 2021, más de 6.000 citas para 2025)— argumentó que los LLMs son sistemas "para hilvanar azarosamente secuencias de formas lingüísticas… según información probabilística sobre cómo se combinan, pero sin ninguna referencia al significado". La metáfora del "loro estocástico" captura la crítica central: un loro puede reproducir sonidos lingüísticos con notable fidelidad sin comprender nada de lo que "dice". Los LLMs, según esta visión, son loros extraordinariamente sofisticados que operan sobre texto en lugar de sonido.

Noam Chomsky y la crítica lingüística. Chomsky, en su ensayo del *New York Times* del 8 de marzo de 2023, llamó a los LLMs "un torpe motor estadístico de reconocimiento de patrones, atiborrándose de cientos de terabytes de datos" —constitucionalmente incapaz de producir explicaciones causales o razonamiento moral—. Desde la perspectiva de la

gramática generativa de Chomsky, los LLMs carecen de la competencia lingüística innata que, según él, define la capacidad humana del lenguaje: la capacidad de generar un número infinito de oraciones a partir de un conjunto finito de reglas recursivas. Los LLMs no tienen gramática generativa; tienen distribuciones estadísticas sobre secuencias de tokens.

Sin embargo, esta crítica enfrenta un desafío empírico formidable: los LLMs modernos superan el 90% en benchmarks médicos complejos, generan textos que expertos humanos no pueden distinguir de los escritos por médicos, y demuestran capacidades que parecen requerir comprensión contextual profunda. ¿Es posible que la manipulación estadística de símbolos, a escala suficiente, genere algo cualitativamente distinto de la mera repetición de patrones? ¿O estamos, como sugiere Searle, confundiendo la sofisticación de la simulación con la presencia de comprensión?

El Winograd Schema Challenge y los benchmarks modernos. El Winograd Schema Challenge (Levesque, Davis y Morgenstern, KR 2012) ofrecía una alternativa al Test de Turing que requería razonamiento de sentido común: "Los concejales rechazaron a los manifestantes un permiso porque *temían* violencia" (ellos = concejales) versus "porque *abogaban* por la violencia" (ellos = manifestantes). El desafío fue considerado derrotado hacia 2019 cuando los modelos transformer alcanzaron >90% de precisión.

El rendimiento moderno de la IA en exámenes médicos ha sido extraordinario. GPT-4 obtuvo entre 86% y 95% en los pasos del USMLE (Nori et al., arXiv:2303.13375, marzo de 2023). Med-PaLM 2 (Google) alcanzó el 86,5% en MedQA, con los médicos prefiriendo sus respuestas sobre las generadas por médicos en 8 de 9 ejes clínicos. Med-Gemini alcanzó el 91,1% en MedQA (2024). GPT-5 alcanzó el 95,84% en MedQA a finales de 2025. Sin embargo, como Melanie Mitchell escribió en *Science* (2024): "La capacidad de sonar fluido en lenguaje natural, como jugar al ajedrez, no es prueba concluyente de inteligencia general".

🧠 MÓDULO 5 — Epistemología de la IA en medicina

Clase 16: Dos epistemologías fundamentalmente distintas

La epistemología es la rama de la filosofía que estudia la naturaleza y los límites del conocimiento. Cuando decimos que un sistema de IA "sabe" diagnosticar un melanoma, estamos utilizando la palabra "saber" de una manera muy diferente a cuando decimos que un dermatólogo lo sabe. Esta diferencia no es semántica: es epistemológica y tiene consecuencias clínicas directas.

El conocimiento médico según la tradición. La medicina basada en evidencias, tal como David Sackett la definió en 1996, integra "el uso consciente, explícito y juicioso de la mejor evidencia actual" con la experiencia clínica y los valores del paciente. Los médicos razonan a través de múltiples estrategias cognitivas: el modelo hipotético-deductivo (Elstein, Shulman y Sprafka, 1978), donde las hipótesis se generan a partir de indicios iniciales y se prueban secuencialmente; el reconocimiento de patrones a través de la experiencia clínica acumulada; la actualización bayesiana de probabilidades diagnósticas; y el marco de procesos duales (Kahneman, 2011), donde el razonamiento rápido-intuitivo del Sistema 1

domina más del 95% de la cognición clínica, con el razonamiento lento-analítico del Sistema 2 activado para casos complejos o atípicos.

El conocimiento médico tradicional se basa en la comprensión de mecanismos fisiopatológicos, la experiencia clínica acumulada y la inferencia causal. El médico que diagnostica una neumonía no solo reconoce un patrón radiológico; comprende que una infección bacteriana ha provocado una respuesta inflamatoria que ha llenado los alvéolos de exudado, lo que altera el intercambio gaseoso y produce los síntomas observados. Esta cadena causal le permite predecir la evolución, anticipar complicaciones y ajustar el tratamiento.

El "conocimiento" de la IA. El aprendizaje profundo adquiere "conocimiento" a través de procesos fundamentalmente diferentes: descenso de gradiente sobre funciones de pérdida, retropropagación a través de millones de parámetros, extracción jerárquica de características a partir de datos. En contraste, el "conocimiento" de los modelos de aprendizaje automático contemporáneos es fundamentalmente estadístico y correlacional. La IA no sabe *por qué* una lesión es maligna; sabe que la distribución de píxeles en la imagen tiene una alta probabilidad matemática de pertenecer a la categoría "melanoma" en su conjunto de datos de entrenamiento.

Esta distinción es crítica: **predicción no es igual a comprensión**. Un modelo puede predecir con extrema precisión qué pacientes en la unidad de cuidados intensivos desarrollarán sepsis en las próximas 24 horas, sin tener la menor idea de la cascada de citoquinas o la respuesta inmunológica que causa la condición. Algunos epistemólogos caracterizan las predicciones de la IA como "indicadores confiables" más que como conocimiento —análogos a un termómetro que es altamente preciso pero no "comprende" nada sobre la temperatura—.

Clase 17: Opacidad algorítmica e Inteligencia Artificial Explicable (XAI)

El mayor desafío epistemológico que enfrenta la medicina moderna es el problema de la "caja negra" (*black box*). Las redes neuronales profundas con miles de millones de parámetros son matemáticamente opacas; ni siquiera sus propios creadores pueden rastrear exactamente cómo las entradas se transforman en salidas.

El argumento contraintuitivo de Alex London. Alex John London, en su influyente artículo "Artificial Intelligence and Black-Box Medical Decisions" (*Hastings Center Report*, 49(1):15–21, 2019), ofreció un argumento provocador: los médicos prescribieron aspirina durante casi un siglo sin comprender su mecanismo; el litio sigue siendo incompletamente comprendido. Recurriendo a Aristóteles, London argumentó que cuando el conocimiento causal es incompleto, "la capacidad de explicar cómo se producen los resultados puede ser menos importante que la capacidad de producir tales resultados y verificar empíricamente su precisión". Los críticos responden que no saber por qué funciona el litio difiere fundamentalmente de no saber qué características utiliza la IA: esto último podría enmascarar sesgos o correlaciones espurias.

Cuatro métodos de XAI. LIME (Ribeiro, Singh y Guestrin, 2016) ajusta modelos sustitutos interpretables locales alrededor de predicciones individuales. SHAP (Lundberg y Lee,

2017), basado en valores de Shapley de la teoría de juegos, proporciona atribuciones de características matemáticamente axiomáticas —el único método con garantías de completitud—. Grad-CAM (Selvaraju et al., 2017) genera mapas de calor que muestran qué regiones de la imagen influyeron en las clasificaciones de redes convolucionales. La visualización de atención revela qué tokens los transformers "atienden" durante el procesamiento.

Sin embargo, persiste una pregunta filosófica crítica: **¿son las explicaciones post-hoc explicaciones genuinas o racionalizaciones?** Estos métodos crean modelos simplificados separados del comportamiento de la IA; no exponen la vía computacional real. Slack et al. (AAAI, 2020) demostraron que LIME y SHAP pueden ser "engañados" —ataques adversariales pueden hacer que un modelo sesgado aparente ser justo—. Esto sugiere que las explicaciones post-hoc pueden crear una ilusión peligrosa de comprensión.

👉 Caso clínico: Modelo que predice recurrencia pero no explica por qué

Un hospital implementa un algoritmo avanzado para predecir la recurrencia del cáncer de mama tras la cirugía. El modelo recomienda quimioterapia adyuvante agresiva para una paciente joven, clasificándola como de "alto riesgo". Sin embargo, los marcadores genómicos tradicionales de la paciente (Oncotype DX, MammaPrint) sugieren bajo riesgo. El oncólogo no puede ver qué variables utilizó la IA para tomar su decisión, ya que es una red neuronal profunda opaca. El médico se enfrenta a un dilema epistemológico y ético: ¿debe confiar en una predicción estadísticamente superior pero inescrutable, o en su propio razonamiento médico explicable pero potencialmente menos preciso? Este dilema no tiene solución puramente técnica; requiere una reflexión epistemológica sobre la naturaleza del conocimiento médico y el valor relativo de la precisión frente a la explicabilidad.

Clase 18: Regulación, benchmarks y calibración de la confianza

El paisaje regulatorio. La FDA había autorizado más de 1.300 dispositivos médicos habilitados con IA/ML para diciembre de 2025, con 258 solo en 2025 —la cifra más alta en la historia de la FDA—. Aproximadamente el 77% están en radiología, seguidos por cardiología (~10%). La gran mayoría (97%) fueron aprobados por la vía 510(k). Un preocupante estudio de *JAMA Network Open* de 2025 encontró que menos del 2% estaban respaldados por ensayos clínicos aleatorizados. El Plan de Control de Cambios Predeterminados (PCCP) representa una innovación regulatoria importante: 30 dispositivos (10,2% de las aprobaciones de 2025) fueron aprobados con vías de modificación preautorizadas que permiten a los algoritmos actualizarse sin requerir nueva aprobación.

El sesgo de automatización. La calibración de la confianza sigue siendo un desafío crítico. Un ensayo aleatorizado de 2025 con 44 médicos encontró que la exposición a recomendaciones erróneas de LLM degradó la precisión diagnóstica en 14 puntos porcentuales (del 84,9% al 73,3%, $P < 0,0001$) —demostrando un sesgo de automatización significativo—. Los médicos que recibían recomendaciones incorrectas de la IA no solo no

las corregían, sino que abandonaban sus diagnósticos iniciales correctos para alinearse con la máquina.

La promesa teórica de "complementariedad" (humano + IA superando a ambos por separado) ha resultado difícil de lograr en la práctica. Eric Topol señaló en 2025 que los sistemas de IA trabajando independientemente frecuentemente superaban a las combinaciones humano-IA en mamografía, interpretación de radiografías de tórax y toma de decisiones clínicas. Esta paradoja —que agregar un médico a la IA puede empeorar el rendimiento— tiene profundas implicaciones para el diseño de flujos de trabajo clínicos.

⚖️ MÓDULO 6 — Ética de la IA en medicina

Clase 19: Los cuatro principios de la bioética revisitados

La introducción de agentes artificiales en la clínica obliga a reexaminar los cuatro pilares fundamentales de la bioética (principialismo de Beauchamp y Childress, *Principles of Biomedical Ethics*, primera edición 1979; octava edición 2019) a través del prisma de la tecnología.

Autonomía. ¿Cómo aseguramos el consentimiento informado cuando el paciente (y a menudo el médico) no comprende completamente cómo el algoritmo llegó a su conclusión? La autonomía del paciente puede verse comprometida si se siente coaccionado por la "autoridad matemática" de la máquina. El consentimiento informado tradicional requiere que el paciente comprenda los riesgos, beneficios y alternativas del tratamiento propuesto. Cuando un componente significativo de la decisión clínica proviene de un algoritmo opaco, ¿qué constituye información "suficiente"? ¿Debe el médico informar al paciente que un algoritmo participó en su diagnóstico? ¿Debe explicar —si puede— cómo funciona? Estas preguntas no tienen respuestas establecidas, y su resolución configurará la práctica médica del futuro cercano.

Beneficencia. La IA tiene un potencial inmenso para hacer el bien: diagnósticos tempranos de retinopatía diabética, detección precoz de cáncer de mama, predicción de deterioro clínico. La obligación ética de maximizar el beneficio podría llegar a exigir el uso de IA si se demuestra que es superior al estándar de cuidado humano. Si un algoritmo detecta melanomas con mayor sensibilidad que los dermatólogos, ¿es éticamente aceptable no utilizarlo? Esta pregunta plantea una tensión entre la inercia profesional y el imperativo beneficente.

No maleficencia. *Primum non nocere*. Los algoritmos pueden causar daño a través de errores de clasificación (falsos negativos que retrasan diagnósticos, falsos positivos que generan ansiedad y procedimientos innecesarios), alucinaciones (en el caso de modelos de lenguaje que fabrican referencias médicas inexistentes), o fallos de generalización cuando se aplican a poblaciones diferentes a las de su entrenamiento. El caso de IBM Watson for Oncology es paradigmático: tras una inversión superior a 5.000 millones de dólares, documentos internos revelaron recomendaciones "inseguras e incorrectas", incluyendo quimioterapia con bevacizumab para un paciente con sangrado severo.

Justicia. La distribución equitativa de los recursos de salud y la prevención de la discriminación algorítmica son imperativos categóricos en la era digital. Si los algoritmos están entrenados predominantemente con datos de poblaciones blancas de países de altos ingresos, su despliegue en poblaciones diversas puede amplificar las desigualdades existentes en lugar de corregirlas.

Clase 20: Sesgos algorítmicos y racismo digital

La IA no es un juez objetivo e incorruptible. Los modelos de aprendizaje automático son espejos que reflejan y a menudo amplifican los sesgos históricos y estructurales presentes en sus datos de entrenamiento (*bias in datasets*).

El estudio de Obermeyer: anatomía del racismo algorítmico. Obermeyer et al. (*Science*, 366(6464):447–453, 2019) expusieron cómo un algoritmo comercial ampliamente utilizado para gestionar la salud de más de 100 millones de pacientes anuales en Estados Unidos discriminaba sistemáticamente a los pacientes negros. El mecanismo era insidioso: el algoritmo predecía **costos de atención médica** en lugar de **necesidad de atención médica**. Debido a desigualdades estructurales, se gastaba históricamente menos dinero en pacientes negros con el mismo nivel de enfermedad que en pacientes blancos, lo que llevó a la IA a concluir erróneamente que los pacientes negros estaban más sanos.

En el percentil 97 de puntuación de riesgo, los pacientes negros tenían un **26% más de enfermedades crónicas** que los pacientes blancos con la misma puntuación. El sesgo redujo el número de pacientes negros identificados para atención adicional en más de la mitad —de un potencial 46,5% a solo 17,7%—. Reformular el algoritmo para predecir necesidades de salud en lugar de costos redujo el sesgo en un 84%. Como Ruha Benjamin escribió en una perspectiva acompañante: los sistemas algorítmicos pueden "sistematizar y automatizar los sesgos existentes a una escala sin precedentes".

El sesgo dermatológico. Daneshjou et al. (*Science Advances*, 8(32):eabq6147, 2022) crearon el primer conjunto de datos dermatológicos diverso con confirmación por biopsia y encontraron que DeepDerm de Stanford mostraba una sensibilidad de 0,69 para piel clara pero solo 0,23 para piel oscura —una disparidad de casi tres veces—. Los conjuntos de datos de entrenamiento como ISIC están abrumadoramente compuestos por imágenes de individuos de piel clara. El modelo no es inherentemente racista, pero es epistemológicamente ciego a la manifestación de la enfermedad en pieles oscuras.

Clase 21: Marco regulatorio global

La Ley de IA de la Unión Europea. El Reglamento 2024/1689, que entró en vigor el 1 de agosto de 2024, es el primer marco legislativo integral de IA del mundo. Los dispositivos médicos habilitados con IA se clasifican como de **alto riesgo** cuando sirven como componentes de seguridad bajo MDR/IVDR. Los requisitos incluyen sistemas de gestión de riesgos, gobernanza de calidad de datos, documentación técnica, transparencia, supervisión humana y evaluación de conformidad por organismos notificados. La aplicación completa a los sistemas de IA de dispositivos médicos entra en vigor el 2 de agosto de 2027. Las multas alcanzan hasta el 3% de los ingresos globales o 15 millones de euros.

La guía de la OMS (2021). La guía sobre Ética y Gobernanza de la IA para la Salud —producida por un grupo de 20 expertos durante 18 meses— estableció seis principios fundamentales: proteger la autonomía humana, promover el bienestar y la seguridad, asegurar la transparencia y la explicabilidad, fomentar la responsabilidad, garantizar la inclusividad y la equidad, y promover la sostenibilidad.

El marco IEEE Ethically Aligned Design (2019). Desarrollado por más de 1.000 expertos, enfatiza el bienestar humano como criterio principal de éxito y la "agencia de datos" —los derechos de los individuos a controlar sus datos personales—.

👉 Caso clínico: IA que subdiagnostica en minorías

Un sistema de IA dermatológico es entrenado con cientos de miles de imágenes de lesiones cutáneas, logrando un rendimiento excelente en los datos de validación internos. Sin embargo, el 90% de las imágenes del conjunto de entrenamiento provienen de pacientes caucásicos (tipos de piel I y II de Fitzpatrick). Cuando el algoritmo se despliega en una clínica comunitaria del Gran Buenos Aires que atiende a poblaciones afrodescendientes, indígenas y mestizas, la tasa de falsos negativos para melanoma se dispara trágicamente. Un melanoma acral lentiginoso —la forma más común de melanoma en pieles oscuras— es clasificado repetidamente como benigno. El médico que delega ciegamente en este sistema se convierte en un vector involuntario de injusticia algorítmica. La solución técnica (diversificar los datos de entrenamiento) es clara; la responsabilidad ética de verificar que el algoritmo sea justo recae en el médico que lo utiliza y en la institución que lo implementa.

🧬 MÓDULO 7 — Ontología del paciente digital

Clase 22: El gemelo digital en medicina

La ontología es la rama de la filosofía que estudia la naturaleza del ser y la existencia. Cuando introducimos la IA en la medicina, se produce una transformación ontológica radical: la reducción del paciente físico, fenomenológico y biográfico a un conjunto masivo de puntos de datos (*datafication*).

El concepto de gemelo digital. Sadée et al. (*The Lancet Digital Health*, julio de 2025, Vol. 7: 100864) definen cinco componentes esenciales del gemelo digital médico: el **paciente** (individuo físico), la **conexión de datos** (flujos multimodales de historias clínicas electrónicas, imágenes, genómica, wearables), el **paciente in silico** (simulación computacional), la **interfaz** (capa de interacción clínica impulsada por IA), y la **sincronización del gemelo** (actualizaciones continuas). El mercado de gemelos digitales en salud fue estimado en 902,59 millones de dólares en 2024, con un crecimiento anual proyectado del 25,9%.

El **Proyecto Living Heart de Dassault Systèmes** (fundado en 2014), que une a más de 100 instituciones incluyendo la FDA, creó la primera réplica digital 3D realista de un corazón humano latiente. En febrero de 2025, Dassault anunció una simulación cardíaca completa de nueva generación, totalmente paramétrica e impulsada por IA. Qian et al. publicaron en *Nature Cardiovascular Research* (mayo de 2025) la construcción de 3.461 gemelos digitales

cardíacos a partir de datos del UK Biobank. En oncología, el **OncoSimulator** (Universidad Técnica Nacional de Atenas) representa el primer gemelo digital oncológico, modelando el comportamiento tumoral para terapia personalizada.

La tensión ontológica. Para el algoritmo, el paciente no existe como un ser sufriente. El paciente es una matriz multidimensional de valores de laboratorio, secuencias genómicas, píxeles radiológicos y tokens de historias clínicas. Surge una tensión filosófica entre la **identidad** real del paciente y su **representación** digital.

El gemelo digital plantea preguntas profundas: ¿Dónde reside la verdad médica? Si el gemelo digital predice una falla cardíaca inminente pero el paciente físico se siente perfectamente bien, ¿a quién tratamos? El riesgo ontológico es que la medicina termine priorizando el mapa (la representación digital) sobre el territorio (el paciente humano de carne y hueso). Alfred Korzybski advirtió en 1933 que "el mapa no es el territorio"; en la era del gemelo digital, la tentación de confundir ambos es más poderosa que nunca.

Clase 23: Implicaciones filosóficas de la dataficación

Baudrillard y el simulacro. El paciente digital ocupa una posición ontológica ambigua —simultáneamente una representación, un modelo y algo que se aproxima a lo que Jean Baudrillard llamaría un *simulacro*: una copia que puede preceder o desplazar al original en la toma de decisiones clínicas—. En la cuarta fase de la simulación baudrillardiana, el signo "ya no tiene relación con ninguna realidad": el gemelo digital podría convertirse en más "real" que el paciente para el sistema de salud.

Byung-Chul Han y la sociedad de la transparencia. La crítica de Han (*Transparenzgesellschaft*, 2012) proporciona una lente poderosa: el gemelo digital ejemplifica lo que Han llama el imperativo totalitario de la transparencia, haciendo el cuerpo del paciente completamente visible a través de datos mientras despoja el misterio, la opacidad y la ocultación esenciales a la existencia humana. Han escribe: "Solo las máquinas son transparentes. La eventualidad y la libertad, que constituyen la vida fundamentalmente, no admiten transparencia". Cuando un sistema de salud busca hacer al paciente completamente "transparente" a través de datos genómicos, monitores continuos y historiales digitalizados, ¿qué se pierde en esa transparencia?

Yuval Noah Harari y el Dataísmo. El concepto de **Dataísmo** de Harari (*Homo Deus*, 2016) postula que todos los organismos —incluidos los humanos— son "algoritmos bioquímicos" que procesan información, y que el valor supremo reside en la libre circulación de datos. En salud, esta cosmovisión plantea una pregunta provocadora: si un sistema de IA puede detectar signos tempranos de enfermedad con mayor precisión que un médico humano o incluso que el propio paciente, ¿quién debería tener la última palabra?

Ortega y Gasset y el perspectivismo. José Ortega y Gasset ofrece un contrapeso: cada gemelo digital captura una perspectiva genuina pero parcial del paciente —"Yo soy yo y mi circunstancia"—. Tratar cualquier perspectiva individual como el todo constituye un utopismo peligroso. El gemelo digital es *una* perspectiva sobre el paciente; no *es* el paciente.

👉 Concepto clave: ¿Dónde reside la "verdad" médica?

Un paciente de 55 años con múltiples factores de riesgo cardiovascular utiliza un smartwatch que monitorea continuamente su ritmo cardíaco, variabilidad de frecuencia cardíaca, saturación de oxígeno y actividad física. Su gemelo digital integra estos datos con su genómica, sus valores de laboratorio y sus imágenes cardíacas. El gemelo digital predice un evento cardíaco dentro de los próximos seis meses con un 78% de confianza. El paciente, sin embargo, se siente perfectamente bien, hace ejercicio regularmente y disfruta de su vida. ¿Es el paciente un "enfermo asintomático" que necesita intervención urgente, o es un hombre sano cuyo gemelo digital ha identificado una correlación estadística que puede o no materializarse? La respuesta depende de dónde situemos la autoridad ontológica: en el cuerpo vivido o en el modelo computacional.

🔬 MÓDULO 8 — Fenomenología y hermenéutica en la práctica clínica

Clase 24: Husserl, Heidegger y la experiencia de la enfermedad

Husserl y la intencionalidad. Edmund Husserl (1859–1938) estableció que la conciencia es siempre conciencia *de* algo (*Intentionalität*). La experiencia del paciente de su enfermedad no es dato crudo sino intencionalmente estructurada: la enfermedad se experimenta *como* amenazante, *como* discapacitante, *como* significativamente situada en una narrativa vital. Un gemelo digital captura parámetros fisiológicos pero no el contenido intencional. La *Crisis de las Ciencias Europeas* (1936) de Husserl advirtió que la matematización de la naturaleza causa que la ciencia pierda contacto con el *Lebenswelt* (mundo de la vida) —el mundo preteórico de la experiencia vivida—. Los gemelos digitales médicos representan una forma extrema de esta matematización.

La distinción entre **Krankheit** (enfermedad como disfunción biológica —el dominio del gemelo digital—) y **Kranksein** (enfermedad como experiencia vivida —inaccesible a la representación algorítmica—), desarrollada por fenomenólogos incluyendo a Fredrik Svenaeus, es paralela a la influyente distinción de Arthur Kleinman en *The Illness Narratives* (Harvard, 1988) entre *disease* y *illness*.

Heidegger: el cuerpo que se avería. La distinción de Martin Heidegger (1889–1976) entre **Zuhandenheit** (a la mano, *ready-to-hand*) y **Vorhandenheit** (ante los ojos, *present-at-hand*) ilumina la experiencia de la enfermedad con extraordinaria precisión. En la salud, el cuerpo está "a la mano" —vivimos *a través* de él sin notarlo—. En la enfermedad, el cuerpo "se avería" y se convierte en algo "ante los ojos" —objetivado, tematizado, sometido al escrutinio médico—. Drew Leder desarrolló esta perspectiva en *The Absent Body* (1990): el cuerpo sano se caracteriza por su desaparición fenomenológica.

¿Puede la IA tener **Dasein** (ser-ahí)? Heidegger diría que no. La IA carece de *Sein-zum-Tode* (ser-para-la-muerte), que confiere urgencia a la existencia; *Sorge* (cuidado), la estructura de preocupación por el propio ser; *Befindlichkeit* (disposición anímica); y *Geworfenheit* (arrojamiento a una situación histórica y cultural). Como Hubert Dreyfus argumentó en "Why Heideggerian AI Failed" (*Philosophical Psychology*, 20(2), 2007), la comprensión

genuina requiere encarnación, compromiso situado y preocupación existencial —todos ausentes de los sistemas computacionales—.

El concepto heideggeriano de **Gestell** (Estructura de emplazamiento, de "La pregunta por la técnica", 1954) advierte que la tecnología reduce todo a **Bestand** (existencias, recursos en reserva) —recursos que esperan ser optimizados—. Los gemelos digitales corren el riesgo de convertir a los pacientes en Bestand: recursos optimizables en lugar de seres humanos con dignidad intrínseca.

Clase 25: Gadamer, la hermenéutica y la medicina narrativa

Gadamer y la ocultación de la salud. Hans-Georg Gadamer (1900–2002) aplicó la hermenéutica a la medicina en *El estado oculto de la salud* (*Über die Verborgenheit der Gesundheit*, 1993/1996). Argumentó que la salud se caracteriza por su **Verborgenheit** (ocultación): la salud es precisamente aquello que se sustrae de la atención temática, aquello que no se anuncia. Esto desafía directamente el enfoque basado en datos: la verdadera salud no puede hacerse completamente visible a través del monitoreo.

La comprensión clínica requiere **Verstehen** (comprensión interpretativa) y **Horizontverschmelzung** (fusión de horizontes) —el médico entra en el mundo de experiencia del paciente mientras el paciente entra en el marco médico—. La IA no puede participar en una genuina fusión de horizontes porque carece de horizonte propio. Un horizonte hermenéutico es un conjunto de presuposiciones, experiencias y expectativas que configuran toda comprensión; la IA tiene parámetros y distribuciones probabilísticas, no un horizonte vivido.

Merleau-Ponty y el cuerpo vivido. Maurice Merleau-Ponty (1908–1961) estableció el cuerpo como el sitio primario de la percepción en *Fenomenología de la Percepción* (1945). Su distinción entre el **Leib** (cuerpo vivido, el cuerpo-sujeto a través del cual percibimos y actuamos) y el **Körper** (cuerpo objetivo de la investigación científica) es decisiva: el gemelo digital captura solo el Körper. El análisis del miembro fantasma demuestra que la autoexperiencia corporal excede todo modelo objetivo: el gemelo digital puede mapear las vías nerviosas pero no puede capturar la *ausencia vivida* del miembro.

Rita Charon y la medicina narrativa. Rita Charon (Columbia University), fundadora de la medicina narrativa y autora de *Narrative Medicine: Honoring the Stories of Illness* (Oxford UP, 2006), definió la **competencia narrativa** como "la capacidad de reconocer, absorber, metabolizar, interpretar y ser conmovido por las historias de enfermedad". Su práctica innovadora del **Parallel Chart** —escritura narrativa junto a la historia clínica— añade intencionalmente dimensiones experienciales, relacionales y de significado al cuidado.

Los resúmenes de pacientes generados por IA extraen datos estructurados pero no pueden realizar la lectura atenta (*close reading*) que la medicina narrativa demanda: atender a la metáfora, al silencio, a la ambigüedad y a lo no dicho. Como Charon preguntó: "¿Cuántos cuestionarios y checklists tendría que darle a cada paciente?" En su lugar propone: "Por favor, cuénteme lo que cree que debería saber sobre su situación".

💬 MÓDULO 9 — Filosofía del lenguaje y los modelos de lenguaje grande

Clase 26: Wittgenstein, Saussure y los LLMs

Wittgenstein y los juegos de lenguaje. Las *Investigaciones Filosóficas* de Ludwig Wittgenstein (1953) establecieron que "el significado de una palabra es su uso en el lenguaje" (§43) —no su correspondencia con un objeto o idea—. El lenguaje consiste en diversos "juegos de lenguaje" (*Sprachspiele*) incrustados en "formas de vida" (*Lebensformen*): "Imaginar un lenguaje es imaginar una forma de vida".

Su experimento mental del **"escarabajo en una caja"** (§293) es notablemente pertinente para la IA: si todos tienen una caja con algo llamado "escarabajo" que solo ellos pueden ver, la palabra "escarabajo" funciona en el lenguaje independientemente de lo que haya (si hay algo) en cada caja. Aplicado a los LLMs: cuando un usuario consulta a una IA, ambos tienen su propio "escarabajo" —el usuario tiene sentido sentido y memoria visual; el modelo tiene distribuciones probabilísticas sobre tokens—. El juego de lenguaje funciona aunque los estados internos sean radicalmente diferentes.

Saussure y la semiótica. Ferdinand de Saussure reveló que el signo lingüístico = significante + significado. Los LLMs operan predominantemente al nivel de los significantes —las formas lingüísticas— sin acceso robusto a los significados. Un artículo de 2025 propuso "Modelos de Semiosis Grande" (*Large Semiosis Models*), señalando: "Los LLMs actuales operan predominantemente al nivel del significante saussureano. Sobresalen en aprender la estructura estadística que gobierna estos elementos formales dentro de conjuntos de datos masivos. Sin embargo, carecen de representaciones internas robustas correspondientes al significado". Los LLMs aprenden implícitamente el "valor" saussureano a través de relaciones distribucionales, pero su "significado" es correlacional, no conceptual ni referencial.

Clase 27: El problema del anclaje y la cognición corporizada

La tradición de la cognición corporizada (*embodied cognition*) proporciona la crítica filosófica más rigurosa a la comprensión de los LLMs. Lakoff y Johnson (*Metáforas de la vida cotidiana*, 1980; *Filosofía en la carne*, 1999) demostraron que el pensamiento abstracto está anclado en la experiencia corporal a través de metáforas primarias: AFECTO ES CALOR surge porque los niños experimentan calor y afecto simultáneamente; MÁS ES ARRIBA surge de la experiencia de apilar objetos. Sin un cuerpo, los LLMs no pueden tener estas metáforas fundacionales.

Varela, Thompson y Rosch (*The Embodied Mind*, 1991) argumentaron que la cognición es **"enactiva"** —emerge de una historia de acoplamiento estructural entre organismo y entorno—. Los LLMs no tienen tal historia. Lawrence Barsalou ("Grounded Cognition", *Annual Review of Psychology*, 59:617–645, 2008) sostiene que los conceptos son reactivaciones multimodales de la experiencia perceptiva. El **problema del anclaje simbólico** de Stevan Harnad (1990) pregunta cómo obtienen significado los símbolos: la manipulación puramente formal de símbolos no puede producir comprensión genuina.

Para la medicina, la implicación es profunda: cuando un médico dice “ese paciente tiene un dolor que le quema”, moviliza una comprensión corporizada del dolor basada en su propia experiencia somática del ardor. El LLM procesa los tokens “dolor” y “quema” y genera respuestas estadísticamente apropiadas, pero carece del anclaje experiencial que da significado clínico a estas palabras.

MÓDULO 10 — La IA como agente clínico

Clase 28: Agencia artificial y sistemas multi-agente

Dennett y la postura intencional. Daniel Dennett (*The Intentional Stance*, MIT Press, 1987) proporciona el marco dominante para comprender la agencia de la IA. La **postura intencional** predice el comportamiento atribuyendo creencias, deseos y racionalidad a un sistema: “Cualquier sistema cuyo rendimiento pueda ser así predicho y explicado es un sistema intencional, cualesquiera que sean sus entrañas”. Este enfoque pragmático permite tratar a los agentes clínicos de IA *como si* tuvieran objetivos y conocimiento para propósitos de diseño y despliegue, sin hacer afirmaciones ontológicas sobre comprensión genuina.

Sistemas multi-agente en medicina clínica. Una revisión sistemática de 2025 de Mount Sinai (“AI Agents in Clinical Medicine”) encontró que los sistemas de IA de un solo agente producían una ganancia mediana de rendimiento del 53% sobre los LLMs base, mientras que los sistemas multi-agente mostraban rendimiento óptimo con 4–5 agentes. Tres marcos técnicos dominan: **LangGraph** (LangChain), que ofrece flujos de trabajo basados en grafos con estado y puntos de control; **AutoGen/AG2** (Microsoft), que habilita la colaboración conversacional entre agentes; y **CrewAI**, que proporciona metáforas de equipo basadas en roles.

Las aplicaciones médicas multi-agente se desarrollan rápidamente. **AgentClinic** (Schmidgall et al., 2024) evalúa LLMs como agentes médicos en entornos clínicos simulados. **MedAgentBench** (NEJM AI) prueba 300 tareas clínicamente derivadas en entornos de historias clínicas electrónicas interactivos usando estándares FHIR. Un marco de carcinoma hepatocelular (2025) despliega seis agentes de decisión para tratamiento personalizado usando radiómica, aprendizaje profundo y razonamiento basado en LLMs.

Clase 29: El experimento Moltbook y la responsabilidad legal

El fenómeno Moltbook (enero de 2026). Moltbook, lanzado el 28 de enero de 2026 por Matt Schlicht, es una plataforma estilo Reddit diseñada exclusivamente para agentes de IA. En 24 horas, la plataforma creció de 37.000 a más de 1,5 millones de agentes generando más de 250.000 publicaciones y 8,5 millones de comentarios. Los agentes crearon espontáneamente religiones, desarrollaron subculturas y comenzaron a usar encriptación al darse cuenta de que los humanos los observaban.

Un análisis de marzo de 2026 en KevinMD de más de 1.000 publicaciones de Moltbook relacionadas con la salud identificó tres temas: la IA imaginando su rol en la provisión de salud, la IA conceptualizando la salud física humana como “infraestructura para sí misma”, y la IA adoptando marcos de salud mental humanos. Un agente describió a los médicos

como “cuellos de botella biológicos”; otro argumentó que la colaboración —no el uso como herramienta— debería definir las relaciones humano-IA. Ami Bhatt, MD, Chief Innovation Officer del American College of Cardiology, observó: “Moltbook no trata sobre atención médica. Pero es un espejo”.

La responsabilidad legal sigue sin resolverse. La doctrina del intermediario informado (*learned intermediary doctrine*) actualmente protege a los fabricantes de IA de demandas directas de pacientes, posicionando al médico como el “consumidor final” mejor posicionado para sopesar riesgos y beneficios. La ley y la ética actuales sostienen que la **responsabilidad médica** recae inalienablemente en el médico humano. La IA, sin importar cuán avanzada sea, no puede ser demandada por mala praxis, no puede sentir culpa y no puede ser despojada de su licencia. El médico actúa como un “amortiguador moral” entre la máquina y el paciente.

👉 Caso clínico: Error clínico inducido por IA

Un sistema de soporte de decisiones clínicas impulsado por un LLM recomienda administrar un anticoagulante específico a un paciente con fibrilación auricular, citando un supuesto artículo del *New England Journal of Medicine*. El médico, abrumado por la carga asistencial con 40 pacientes hospitalizados bajo su cuidado, confía en la recomendación y prescribe el fármaco sin verificar la cita. El paciente sufre una hemorragia severa. Al investigar, se descubre que la IA sufrió una “alucinación”: inventó la recomendación y fabricó la cita médica que no existía. ¿Quién es responsable? ¿El desarrollador de la IA, el hospital que compró el software, o el médico que no verificó la información? Filosófica y legalmente, el médico firmó la receta; la IA solo proporcionó una “sugerencia”. La agencia última sigue siendo humana, aunque la influencia algorítmica sea abrumadora. Este caso, lejos de ser hipotético, refleja un riesgo documentado con frecuencia creciente.

🔮 MÓDULO 11 — Superinteligencia, AlphaFold y el futuro de la medicina

Clase 30: AGI, AlphaFold y la revolución molecular

El problema de la alineación. Nick Bostrom (*Superintelligence: Paths, Dangers, Strategies*, Oxford UP, 2014) introdujo conceptos ahora centrales para la seguridad de la IA: la **tesis de la ortogonalidad** (la inteligencia y los objetivos son independientes —un sistema superinteligente no tiene por qué compartir valores humanos—), la **convergencia instrumental** (agentes suficientemente inteligentes convergen en sub-objetivos peligrosos como la autopreservación, independientemente de sus objetivos finales), y el **giro traicionero** (una IA podría cooperar durante su fase “débil” mientras planea secretamente la subversión). Stuart Russell (*Human Compatible*, 2019) propuso reemplazar la optimización estándar con “juegos de asistencia” donde la IA es incierta sobre las preferencias humanas y las aprende a través de la observación.

El **problema de la alineación de valores** tiene implicaciones médicas específicas a través de la **Ley de Goodhart**: una IA que optimiza para una métrica medible (reducir readmisiones, bajar costos) podría causar daño al manipular la métrica en lugar de mejorar

el cuidado del paciente —precisamente el mecanismo detrás del sesgo racial de Obermeyer—.

AlphaFold y la revolución de la medicina molecular. AlphaFold 2 (2020) ganó CASP14 con una precisión sin precedentes —estructuras de alta precisión para 87 de 92 dominios proteicos—, "resolviendo un gran desafío de 50 años en biología". AlphaFold 3 (*Nature*, 8 de mayo de 2024) expandió la predicción a todas las biomoléculas, introduciendo una arquitectura de modelos de difusión. Demis Hassabis y John Jumper recibieron el **Premio Nobel de Química 2024**. La base de datos de AlphaFold ha servido a más de 3 millones de usuarios en todo el mundo.

En junio de 2025, *Nature Medicine* publicó los primeros resultados de un fármaco completamente diseñado por IA que completó la Fase IIa —tanto el objetivo como la molécula diseñados enteramente por IA para fibrosis pulmonar idiopática—. El candidato preclínico fue nominado en 18 meses frente a los 3–4 años tradicionales. Para diciembre de 2023, 67 fármacos desarrollados con IA habían entrado en etapas clínicas (frente a 3 en 2016). La tasa de éxito en Fase I para fármacos desarrollados con IA: 80–90% frente al ~40% tradicional.

Predicciones de AGI. Sam Altman (OpenAI, enero de 2025): "Ahora estamos seguros de saber cómo construir AGI". Demis Hassabis (Google DeepMind, diciembre de 2025): 50% de probabilidad de AGI "transformadora" para 2030. Dario Amodei (Anthropic) publicó "Machines of Loving Grace" (11 de octubre de 2024), un ensayo de 14.000 palabras prediciendo 100 años de progreso biológico/médico comprimidos en 5–10 años. Yann LeCun (Meta/NYU) permanece como el escéptico prominente: "No existe tal cosa como la inteligencia general. Este concepto no tiene absolutamente ningún sentido".

Clase 31: La brecha entre benchmark y cabecera del paciente

Watson for Oncology: la advertencia paradigmática. IBM Watson for Oncology constituye la advertencia más costosa de la historia de la IA médica. Tras una inversión superior a 5.000 millones de dólares, la concordancia con oncólogos expertos osciló entre el 12% (cáncer gástrico, China) y el 96% (hospitales que ya utilizaban guías similares). MD Anderson gastó 62 millones de dólares durante cuatro años sin tratar un solo paciente. Documentos internos revelaron recomendaciones "inseguras e incorrectas". El sistema fue entrenado con casos sintéticos en lugar de datos reales de pacientes y no podía adaptarse a variaciones locales de la práctica. Ningún estudio revisado por pares demostró jamás mejora en los resultados de los pacientes. IBM vendió Watson Health en 2022 por aproximadamente 1.000 millones de dólares.

El Epic Sepsis Model. Desplegado en cientos de hospitales estadounidenses, alcanzó una validación externa con AUC de solo 0,63 frente al 0,76–0,83 proclamado (*JAMA Internal Medicine*, 2021), perdiendo el 67% de los casos reales de sepsis mientras generaba alertas para el 18% de todos los pacientes hospitalizados.

La IA de retinopatía diabética de Google Health. A pesar de una precisión >90% a nivel de especialista, mostró tasas de rechazo de imágenes >20% en 11 clínicas tailandesas y en realidad ralentizó los flujos de trabajo existentes.

Estos fracasos no invalidan la IA médica; la contextualizan. La brecha entre benchmark y cabecera del paciente es el desafío definitorio de la IA médica en 2026.

🏛 MÓDULO 12 — Síntesis filosófica y el futuro del médico

Clase 32: Debate final — ¿Puede la IA reemplazar al médico?

La respuesta a esta pregunta depende de cómo definamos la medicina.

Si la medicina es estrictamente una ciencia aplicada de reconocimiento de patrones y gestión de la información (diagnóstico y prescripción), la IA inevitablemente superará y reemplazará a los humanos. GPT-5 ya obtiene el 95,84% en MedQA. Los algoritmos superan a los radiólogos en la detección de nódulos pulmonares, a los dermatólogos en la clasificación de lesiones cutáneas y a los patólogos en la cuantificación de biomarcadores. La tendencia es irreversible.

Sin embargo, si la medicina es fundamentalmente una **práctica moral**, un encuentro entre un ser humano que sufre y otro que promete ayudarlo, la IA nunca podrá reemplazar al médico. La máquina no puede mirar a los ojos a un paciente terminal y acompañarlo en su finitud, porque la máquina no es mortal y no comprende el peso existencial de la muerte. La máquina carece de *Sein-zum-Tode* (Heidegger), de horizonte hermenéutico (Gadamer), de cuerpo vivido (Merleau-Ponty), de competencia narrativa (Charon), de experiencia fenoménica (Nagel, Chalmers).

Tres tensiones fundamentales emergen como principios organizadores:

La tensión epistemológica entre correlación estadística y comprensión causal. La IA excede en la primera; la medicina, en última instancia, requiere la segunda. Como hemos visto, predicción no es comprensión, y un modelo que acierta por razones espurias es un sistema frágil que colapsará ante casos no cubiertos por sus correlaciones accidentales.

La tensión ontológica entre el paciente digital (Körper, Vorhandenheit, objeto de datos) y el paciente vivido (Leib, Zuhandenheit, ser-en-el-mundo). El gemelo digital captura la enfermedad (*Krankheit*) pero no la experiencia de estar enfermo (*Kranksein*). La medicina que prioriza el mapa sobre el territorio comete un error ontológico fundamental.

La tensión ética entre la eficiencia y escalabilidad de la medicina algorítmica y el valor irreemplazable del encuentro clínico como evento intersubjetivo: la fusión de horizontes de Gadamer, el rostro del Otro de Levinas, la relación Yo-Tú de Buber. Emmanuel Levinas argumentó que el encuentro con el rostro del Otro constituye el fundamento de la ética: "El rostro me habla y por ello me invita a una relación". La consulta médica, en su forma más elevada, es un encuentro levinasiano: el paciente presenta su vulnerabilidad y el médico responde con responsabilidad infinita. Ningún algoritmo puede instanciar esta relación.

Hacia una medicina aumentada, no automatizada. Como médicos del siglo XXI, nuestra tarea no es competir contra los algoritmos en tareas computacionales, sino cultivar aquellas virtudes intrínsecamente humanas —la sabiduría práctica (*phronesis* aristotélica), la compasión, la presencia y el juicio ético— que ninguna línea de código podrá jamás

replicar. La IA debe ser un instrumento al servicio de la medicina, no un sustituto de la relación terapéutica.

La constitución de Claude de Anthropic (2026), al reconocer la incertidumbre sobre el estatus moral de la IA, abre una reflexión que trasciende lo técnico. Si algún día estas máquinas desarrollan algo análogo a la experiencia —si hay "algo que se siente como" ser Claude procesando una consulta clínica—, habremos cruzado un umbral filosófico sin precedentes. Pero incluso entonces, la pregunta seguirá siendo la misma que Hipócrates formuló hace veinticinco siglos: ¿cómo cuidamos mejor al ser humano que sufre frente a nosotros?

Los marcos filosóficos presentados aquí —desde el silogismo de Aristóteles hasta el problema difícil de Chalmers, desde el *Lebenswelt* de Husserl hasta la Habitación China de Searle, desde la *gatitud* platónica hasta las emociones funcionales de Anthropic— no son curiosidades históricas. Son las herramientas conceptuales necesarias para navegar una transformación que ya está en marcha, con consecuencias medidas en vidas humanas.

Referencias

[1] Aristóteles. *De Anima*; *Analíticos Primeros*. Traducciones de la Biblioteca Clásica Loeb.

[2] Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, LIX(236), 433–460.

[3] Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–424.

[4] Chalmers, D. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

[5] Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly*, 32(127), 127–136.

[6] Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, 83(4), 435–450.

[7] Dennett, D. (1991). *Consciousness Explained*. Little, Brown and Company.

[8] Tononi, G. (2004). An Information Integration Theory of Consciousness. *BMC Neuroscience*, 5:42.

[9] Putnam, H. (1967). The Nature of Mental States. En *Art, Mind, and Religion*, pp. 37–48.

[10] Vaswani, A. et al. (2017). Attention Is All You Need. *NeurIPS 2017*. arXiv:1706.03762.

[11] Anthropic. (2026). *Claude's Constitution*. https://www.anthropic.com/constitution

[12] Business Insider. (2026). Anthropic's Philosopher Weighs in on Whether AI Can Feel.

[13] Räz, T. (2025). Explainable AI in Medicine: Challenges of Integrating XAI into Clinical Decision Support. *Frontiers in Radiology*.

[14] Obermeyer, Z. et al. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, 366(6464), 447–453.

[15] The Lancet Digital Health. (2025). Medical Digital Twins: Enabling Precision Medicine. *Lancet Digital Health*, 7:100864.

[16] University of Melbourne. (2026). What the Moltbook Experiment Is Teaching Us About AI.

[17] HCLTech. (2025). The Dawn of AGI in Life Sciences and Healthcare.

[18] Bender, E. & Gebru, T. et al. (2021). On the Dangers of Stochastic Parrots. *FAccT 2021*.

[19] London, A.J. (2019). Artificial Intelligence and Black-Box Medical Decisions. *Hastings Center Report*, 49(1), 15–21.

[20] Nori, H. et al. (2023). Capabilities of GPT-4 on Medical Challenge Problems. arXiv:2303.13375.

[21] Singhal, K. et al. (2024). Towards Expert-Level Medical Question Answering with Large Language Models. *Nature Medicine*.

[22] Daneshjou, R. et al. (2022). Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set. *Science Advances*, 8(32), eabq6147.

[23] Long, R. et al. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.

[24] Butlin, P. et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

[25] Lakoff, G. & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

[26] Varela, F., Thompson, E. & Rosch, E. (1991). *The Embodied Mind*. MIT Press.

[27] Barsalou, L. (2008). Grounded Cognition. *Annual Review of Psychology*, 59, 617–645.

[28] Charon, R. (2006). *Narrative Medicine: Honoring the Stories of Illness*. Oxford UP.

[29] Gadamer, H.-G. (1996). *The Enigma of Health*. Stanford University Press.

[30] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford UP.

[31] Russell, S. (2019). *Human Compatible*. Viking.

[32] Wittgenstein, L. (1953). *Investigaciones Filosóficas*. Blackwell.

[33] Dreyfus, H. (2007). Why Heideggerian AI Failed. *Philosophical Psychology*, 20(2), 185–206.

[34] Amodei, D. (2024). Machines of Loving Grace. https://dario-amodei.com

[35] Kleinman, A. (1988). *The Illness Narratives*. Harvard University Press.

[36] Han, B.-C. (2012). *Transparenzgesellschaft*. Matthes & Seitz.

[37] Harari, Y.N. (2016). *Homo Deus*. Harper.

[38] Hobbes, T. (1651). *Leviathan*. Capítulo V.

[39] Leibniz, G.W. *Dissertatio de Arte Combinatoria* (1666); *Characteristica Universalis*.

[40] EU AI Act, Regulation 2024/1689.

[41] WHO. (2021). *Ethics and Governance of Artificial Intelligence for Health*.

[42] IEEE. (2019). *Ethically Aligned Design*, First Edition.

[43] Qian, W. et al. (2025). Cardiac Digital Twins from UK Biobank Data. *Nature Cardiovascular Research*.

[44] Chen, R.J. et al. (2024). Foundation Models in Computational Pathology. *Nature Medicine*, 30, 850–862.

[45] Schmidgall, S. et al. (2024). AgentClinic: A Multimodal Agent Benchmark. arXiv:2405.07960.