1

INTRODUCCIÓN

La ética de la inteligencia artificial constituye un campo multidisciplinario crítico que examina las implicaciones morales de los sistemas autónomos a lo largo de todo su ciclo de vida, desde el desarrollo inicial hasta la implementación y gobernanza generalizadas (Siau & Wang, 2020; Usher & Barak, 2024). A pesar de que el concepto de «ética de las máquinas» se propuso alrededor de 2006, la ética de la IA sigue en su infancia (Siau & Wang, 2020). El campo es doble: la ética de la IA se ocupa de los principios, reglas, directrices, políticas y regulaciones éticos que rodean a la IA, mientras que la IA ética se ocupa del diseño de sistemas que de hecho se comportan de manera ética (Siau & Wang, 2020). Lo que está en juego es considerable, pues la IA sustenta hoy aplicaciones de alto riesgo —como el reconocimiento facial, la práctica sanitaria y los automóviles autónomos— en las que una decisión equivocada puede costar vidas (Mehta et al., 2023). En respuesta, los gobiernos, la industria y la academia han publicado un extenso panorama de directrices éticas; sin embargo, emergen divergencias significativas dentro de estos documentos y entre ellos, y persiste la incertidumbre sobre cómo deberían implementarse sus principios (Jobin et al., 2019). Esta revisión examina los problemas éticos fundamentales que surgen del desarrollo y el uso de la inteligencia artificial mediante la cartografía de los principales problemas no resueltos que continúan resistiendo toda resolución.

1.1 DEFINICIÓN FORMAL DE LA INTELIGENCIA ARTIFICIAL

Antes de que pueda proceder el análisis ético, el objeto de análisis debe definirse con cierta precisión. La inteligencia artificial se entiende mejor, a los efectos de esta revisión, como el estudio

y la construcción de sistemas artificiales que perciben entornos, procesan información, razonan o infieren, toman decisiones y actúan en pos de objetivos (Schoen, 2023). Un sistema inteligente queda así caracterizado funcionalmente —como un agente que exhibe un comportamiento racional en relación con su entorno y sus objetivos— y no por ninguna afirmación sobre sus estados internos (Devedžić, 2022; Schoen, 2023). Esta definición es deliberadamente agnóstica respecto del sustrato: los sistemas pertinentes se realizan típicamente como componentes de IA incrustados en sistemas de software y hardware más amplios, donde interactúan con datos, modelos y usuarios (Martínez-Fernández et al., 2022). Sus competencias características incluyen la percepción mecánica, el razonamiento, la interpretación y la generación de resultados que orientan la acción (Barenkamp et al., 2020).

Son necesarios dos refinamientos. Primero, la definición debe permanecer cautelosa respecto de la conciencia y el pensamiento genuino: la simulación conductual de resultados inteligentes no establece una comprensión semejante a la humana, y la distinción entre IA estrecha —sistemas competentes en un dominio restringido— e IA general, que igualaría o superaría la competencia humana en todos los dominios, sigue siendo central para calibrar las afirmaciones éticas (Zawacki-Richter et al., 2019). Segundo, la caracterización funcional conecta naturalmente con la tradición formal de la representación del conocimiento, en la cual la actividad inteligente se analiza en términos de agentes, objetivos, observaciones, acciones, recompensas, símbolos y reglas de inferencia (Devedžić, 2022; Lee, 2024; Sowa, 1999). Bajo esta visión, un sistema de IA es un dispositivo que manipula símbolos formales que representan proposiciones, derivando nuevas representaciones según procedimientos definidos (Lee, 2024). La definición operativa adoptada a lo largo de esta revisión es, por tanto, la siguiente: la inteligencia artificial comprende sistemas diseñados que perciben, representan, razonan, deciden y actúan —sin presuponer conciencia, y con el alcance y los límites de sus operaciones formales siempre abiertos al escrutinio.

1.2 LA INTUICIÓN COMO FRONTERA DE LA INTELIGENCIA DE LAS MÁQUINAS

Si la IA se define funcionalmente como percepción, representación, razonamiento y acción, surge la cuestión de si la intuición humana marca una frontera que los sistemas diseñados no pueden cruzar. La literatura filosófica sugiere un punto de partida deflacionario: la intuición no debe tratarse como una facultad misteriosa, sino que puede entenderse como razonamiento discursivo ordinario combinado con una observación adecuada —una competencia argumentativa en la que el razonador extrae conclusiones a partir de consideraciones ya disponibles para él (Servín, s.f.-a). Bajo esta lectura, que Hernández Servín toma de Hintikka, la intuición se dirige a los particulares más que a las verdades generales, y las afirmaciones generales requieren inferencia inductiva o abductiva más allá de lo que la intuición suministra (Servín, s.f.-a). La intuición es, por tanto, un *punto de partida* que debe reconstruirse como argumento: premisas explícitas, inferencias válidas, interpretación de los términos y examen de contraejemplos. «Una demostración intuitiva no es conocimiento», mientras que la demostración de una proposición, aunque sea intuitivamente obvia, sí lo es (Hernández Servín, 2026).

El debate entre Herbert Simon y Hubert Dreyfus adquiere entonces una forma precisa. Simon trató la intuición como reconocimiento de patrones basado en la experiencia —reconocimiento rápido de estructuras familiares que las máquinas pueden en principio simular—, mientras que Dreyfus sostuvo que la intuición humana es una pericia encarnada y sensible al contexto (Gobet & Chassy, 2008). Conviene leer a Dreyfus de primera mano, porque su posición es más fina que una negación global. En *What Computers Can't Do* el núcleo del argumento es la *relevancia*: si toda conducta siguiera reglas, harían falta reglas para aplicar las reglas, y la relevancia de un elemento depende de un contexto que solo puede reconocerse desde un contexto más amplio (el cuchillo significa algo distinto en la cocina, en el quirófano o en una pelea); los seres humanos están «ya en una situación» en la que los hechos llegan interpretados, y «los computadores solo pueden tratar con hechos, pero el hombre, fuente de los hechos,

no es un hecho» (Dreyfus, 1972, pp. 198-203). En su modelo de cinco etapas de la adquisición de habilidades, el experto ya no sigue reglas: ve la situación y actúa intuitivamente (Dreyfus & Dreyfus, 1986, pp. 115-121). Y en 1987 el propio Dreyfus vio en el conexionismo una «nueva esperanza»: los sistemas distribuidos en paralelo reconocen semejanzas «sin usar inferencias ni rasgos aislados», de modo que los computadores «quizá algún día exhiban habilidad» (Dreyfus, 1987, pp. 30-31). Dreyfus separa así su crítica a la IA simbólica de una imposibilidad general, y su pregunta sigue abierta en otra forma: si la relevancia *aprendida* a partir de datos es relevancia *para alguien en una situación*.

La investigación cognitiva contemporánea apoya una posición intermedia: la intuición implica un reconocimiento de patrones rápido y a menudo inexplicado que se apoya en conocimiento implícito adquirido mediante la experiencia (Duch, 2007), y los sistemas híbridos que combinan búsqueda deliberativa con aprendizaje sin modelo se asemejan a la combinación humana de razonamiento e intuición (Vidulich & Tsang, 2026). Duch caracteriza operativamente la intuición por dos rasgos —la solución se alcanza rápidamente y no puede explicarse por qué se eligieron los pasos— y distingue la sensación de solución de la comprensión, que solo llega «cuando todos los pasos intermedios están correctamente enlazados» (Duch, 2007). Las salidas de un modelo estadístico que decide por semejanza tienen esa misma forma. Pueden llamarse, en este sentido preciso, *intuiciones artificiales*: juicios rápidos, por semejanza y no articulables paso a paso, potencialmente valiosos, pero no autojustificados. Esta noción no reduce la intuición humana a reconocimiento de patrones ni declara resoluble la opacidad; admite, con Duch, que una máquina puede producir juicios de tipo intuitivo y niega, con Dreyfus, que esa producción constituya por sí sola justificación o comprensión situada.

Para la ética de la IA, las consecuencias pertinentes son prácticas. Si la intuición de las máquinas se entiende como reconocimiento de patrones sobre representaciones aprendidas, entonces hereda todos los límites epistémicos discutidos más adelante en esta revisión: sus resultados son tan confiables como las representaciones y los datos de los que se derivan, y su carácter rápido y opaco plantea problemas de explicabilidad y sesgo (Afrikanov,

2025). Como toda intuición, una salida de IA vale como *base* para un argumento y no como su conclusión: aceptarla exige que alguien la reconstruya, la contraste y responda por ella. La frontera entre el juicio humano y el análisis algorítmico —quién revisa qué, y con qué autoridad— se convierte así en una cuestión de diseño ético por derecho propio (Afrikanov, 2025). Dreyfus ofrece aquí un puente con la rendición de cuentas que se discute más adelante: en su modelo, el novato que aplica reglas no se siente responsable del resultado, mientras que quien elige una perspectiva se implica en él, y esa implicación es la que fija la experiencia de la que surge la pericia (Dreyfus & Dreyfus, 1986, pp. 117-118). Pericia y responsabilidad aparecen juntas.

Cabe una tesis más fuerte, que esta revisión registra pero no necesita: que el juicio moral exige corporalidad y vulnerabilidad existencial, y que ninguna arquitectura podrá poseerlas. El argumento contra la delegación del juicio no depende de esa predicción sobre las arquitecturas futuras. Basta con que la aceptación de una salida sea un acto de alguien que debe poder dar razones, y con que la responsabilidad por las decisiones recaiga en personas e instituciones. La intuición funciona, por tanto, en esta revisión no como una contrafuerza mística frente a la IA, sino como un concepto fronterizo preciso: nombra las competencias humanas —el juicio sobre los particulares, la pericia encarnada y la rendición de cuentas por las decisiones— que los sistemas algorítmicos pueden imitar pero no certificar.

1.3 EL HEXÁGONO LÓGICO-SEMIÓTICO: UN INSTRUMENTO PARA EL ESTATUS DE LAS AFIRMACIONES Y LOS USOS DE LA IA

La definición de inteligencia artificial puede ampliarse por una vía histórica y estructural que va de los dispositivos mecánicos medievales de razonamiento a los marcos formales-semióticos contemporáneos. Sowa relata que el *Ars Magna* de Ramón Llull comprendía discos con círculos concéntricos giratorios —hechos de metal o pergamino— cuyos sectores, codificados por letras del alfabeto que remitían a atributos y sus adjetivos correspondientes (nueve atributos de Dios, por ejemplo, alineados con sus adjetivos), podían rotarse para generar todas las combinacio-

nes posibles de los atributos (Sowa, 1999). De este modo, Llull construyó discos no solo para la teología, sino también para la filosofía, el derecho, la medicina y los cuatro elementos, y presentó cien sermones basados en combinaciones generadas con sus círculos giratorios (Sowa, 1999). Al girar los círculos, Llull consultaba tablas, árboles y diagramas suplementarios derivados de las categorías y los silogismos de Aristóteles para comprobar si cada combinación era lógicamente permisible —anticipando en efecto el algoritmo de IA más básico: generar y probar (*generate and test*) (Sowa, 1999). Su disco mayor, la *Figura Universalis*, tenía 14 círculos concéntricos con 16 sectores cada uno, un espacio exponencial de más de 72 cuatrillones de combinaciones —una de las primeras explosiones combinatorias registradas en la historia (Sowa, 1999). Sowa concluye que los diagramas y discos del *Ars Magna* constituyen el primer intento de desarrollar ayudas mecánicas al razonamiento, y que sirvieron de inspiración a Gottfried Wilhelm Leibniz (Sowa, 1999).

A través de Leibniz, esta línea combinatoria se vuelve fundacional para la IA simbólica. La concepción de que el razonamiento es una forma de cálculo —no distinta de la aritmética, pero ejecutada sobre símbolos que representan proposiciones en lugar de números— fue formulada por primera vez por Leibniz en el siglo XVII (Lee, 2024). La continuidad es estructural además de histórica: Sowa señala la sorprendente similitud entre el formalismo moderno de representación del conocimiento KL-ONE y el medieval Árbol de Porfirio, correspondiendo las flechas de doble línea del primero a las relaciones de especie a género del segundo (Sowa, 1999).

El hexágono lógico-semiótico pertenece a la misma estirpe de figuras de oposición. Como objeto geométrico, el hexágono es una figura de seis vértices, históricamente etiquetados A-U-E-O-Y-I (Servín, s.f.-b); como objeto lógico, es el hexágono de Blanché, entendido como una extensión del cuadrado semiótico de Greimas, con el cuadrado lógico implícito en él, y se inscribe en una tradición de figuras de oposición que se remonta a Aristóteles, Apuleyo y Boecio (Servín, s.f.-b). El relato histórico de Moretti rastrea el cuadrado lógico desde su origen con Apuleyo y Boecio sobre la base de la lógica silogística aristotélica, pasando por el análisis newtoniano de la descomposición de la luz y el círculo

de colores de Goethe, hasta el modelado por Piaget de las operaciones psico-motoras básicas de la inteligencia humana y de la noción de «estructura», y la dualidad matemática de Gottschalk, que a su vez inspiró a los estructuralistas Jakobson, Lévi-Strauss y Lacan (Servín, s.f.-b).

La tesis de Hernández Servín es que existe un formalismo estructural —en el discurso literario no menos que en las matemáticas— cuyo fundamento es el hexágono lógico-semiótico, arraigado en la teoría de categorías y que posee una lógica interna no necesariamente booleana y dependiente del modelo que describen las categorías en cuestión; la evidencia citada incluye la literatura, el número y la lógica modal, todos los cuales exhiben una estructura visualizable mediante el hexágono (Servín, s.f.-b). Debido a que los operadores de obligación y permisividad pueden representarse formalmente en términos de los operadores modales $\Box$ y $\Diamond$, un hexágono de oposición que relaciona conceptos modales produce por extensión un hexágono para la modalidad deóntica (Servín, s.f.-b); bajo esta lectura, el hexágono lógico-semiótico produce una convergencia entre el lenguaje, el texto, la semiótica y la lógica modal kripkeana (Servín, s.f.-b).

Esta genealogía no debe leerse como una definición de la IA ni como una descripción de cómo funcionan internamente los sistemas actuales. Los modelos de aprendizaje profundo operan sobre espacios latentes continuos que no obedecen a categorías de oposición discretas, y el hexágono no explica ni reduce su opacidad. Lo que la genealogía muestra es otra cosa: que la IA no se agota en el aprendizaje estadístico —la manipulación formal de proposiciones es una de sus líneas fundacionales (Lee, 2024)— y que las figuras de oposición ofrecen un instrumento para un problema que sí es central en la ética de la IA: clasificar con precisión el *estatus* de lo que se afirma y se decide con ayuda de un sistema. La distinción pertinente es la de Newell entre el nivel de conocimiento, donde se analiza qué afirma un sistema, con qué semántica y qué se sigue de ello, y el nivel simbólico, donde se analizan su arquitectura y sus algoritmos (Brachman & Levesque, 2004, pp. 11-12). El hexágono opera en el primer nivel y, por eso, se aplica a las salidas de cualquier sistema, simbólico,

conexionista o híbrido, del mismo modo que la distinción entre validez y solidez que se examina al final de esta revisión.

Siguiendo la extensión deóntica que el propio Hernández Servín sugiere (Servín, s.f.-b), pueden construirse dos hexágonos de Blanché. En ambos, las relaciones son: contradictorias A–O, E–I, U–Y; contrarias A–E, E–Y, Y–A; subcontrarias I–O, O–U, U–I; y subalternas A→I, A→U, E→O, E→U, Y→I, Y→O.

Vértice	Hexágono deóntico (uso p de IA)	Hexágono epistémico (afirmación p)
A	Op: obligatorio	$\Box p$: establecida por la evidencia
E	O$\neg p$: prohibido	$\Box \neg p$: descartada
I	Pp: permitido	$\Diamond p$: compatible con la evidencia
O	P$\neg p$: omisible	$\Diamond \neg p$: no establecida
U	O$p \vee$ O$\neg p$: determinado (la norma debe decidirlo)	$\Box p \vee \Box \neg p$: cuestión decidida
Y	P$p \wedge$ P$\neg p$: facultativo	$\Diamond p \wedge \Diamond \neg p$: cuestión abierta

El hexágono deóntico corresponde a la lógica deóntica del sistema OS (O$p \equiv \neg$P$\neg p$), con semántica de Kripke, que sostiene lenguajes de agentes como IMPACT y que Vidal-López, González-Ruiz, Marcial-Romero y Hernández-Servín proponen reinterpretar en teoría de topos para agentes jurídicos BDI (Vidal-López et al., 2022, pp. 84-88). No es, por tanto, una pieza arqueológica, sino la lógica de los sistemas normativos multiagente actuales.

Su utilidad es prescriptiva en tres sentidos. Primero, detecta *contradicciones*: si un lineamiento universitario establece que «el docente debe revisar toda calificación generada con IA» (A) y otro que «en grupos numerosos puede omitirse la revisión» (O), ambos son contradictorios salvo que se precise el ámbito de cada uno. Segundo, detecta *omisiones*: una política que deja un uso de alto impacto como facultativo (Y) cuando debería estar determinado (U) tiene un vacío normativo. Tercero, detecta *so-*

breafirmaciones: una predicción de «riesgo alto» de reincidencia está, como mucho, en I —compatible con la evidencia estadística sobre un grupo— y se usa como si estuviera en A respecto de un individuo; una salida fluida de un modelo de lenguaje sin anclaje en fuentes ni reconstrucción argumentativa está, como máximo, en I. Nada de esto exige abrir la caja negra: lo exigible es una explicación de razones y un estatus declarado, no necesariamente el mecanismo.

Tres advertencias acotan el instrumento. La primera concierne a la negación: ambos hexágonos presuponen negación clásica, pero la lógica interna del hexágono no es necesariamente booleana sino dependiente del modelo (Servín, s.f.-b); con información incompleta la lógica adecuada puede ser intuicionista —y entonces «no establecida» no equivale a «descartada»—, y con fuentes inconsistentes puede ser paraconsistente, lo que altera las relaciones de oposición. La segunda concierne al coste: toda formalización paga su expresividad con tratabilidad, y ampliaciones pequeñas de un lenguaje pueden volver intratable la detección de inconsistencias (Brachman & Levesque, 1984; Brachman & Levesque, 2004, p. 327). La tercera concierne a la existencia: el hexágono no garantiza que las categorías elegidas sean correctas, exhaustivas o culturalmente adecuadas, y una teoría que lo aplique a la IA necesita al menos un modelo concreto, poblado con políticas reales y evaluable (Servín, s.f.-b). Su formalización categorial y su aplicación a lineamientos institucionales concretos quedan como agenda de investigación.