Ethics in AI

Joseph Servin

September 2026

ÍNDICE GENERAL

I Ética en la Inteligencia Artificial

1 Introducción 9
- 1.1 Definición formal de la inteligencia artificial 9
- 1.2 La intuición como frontera de la inteligencia de las máquinas 11
- 1.3 El hexágono lógico-semiótico: un instrumento para el estatus de las afirmaciones y los usos de la IA 13

2 Revisión de la literatura 19
- 2.1 El problema de la alineación de valores 19
- 2.2 Equidad y sesgo 20
- 2.3 Explicabilidad, transparencia y rendición de cuentas 21
- 2.4 Responsabilidad y rendición de cuentas en el uso cotidiano de la IA 22
- 2.5 Moral de las máquinas en decisiones de alto riesgo 24
- 2.6 Gobernanza, regulación y ejecución 25
- 2.7 La metacuestión: ¿puede resolverse la ética de la IA? 26
- 2.8 El dilema estabilidad-plasticidad: aprendizaje continuo y generalidad de la IA 27
- 2.9 Crítica radical contra la tecnología: el desafío de Kaczynski 30
- 2.10 Los límites epistémicos de la representación: por qué las afirmaciones de «cualquier problema» son injustificadas 35
- 2.11 Validez, interpretación y la ética de la autoridad epistémica delegada 37

3 Metodología 41

4 Resultados 43

5 Discusión 45

6 Conclusión 47

Bibliografía 49

[Editorial note: the definitional passage drafted in earlier sessions is not present in the current document file — insert the existing draft under this heading.]

Aquí está la traducción completa al español, Joseph. Traduje el manuscrito corregido en su totalidad (incluidas las secciones de la definición formal y de la intuición, ya integradas), conservando todas las citas [cite:N#E] y todas las cifras exactamente; excluí únicamente los fragmentos de comentario editorial que contaminaban el archivo (notas de asistente, avisos de colocación y bloques duplicados).

Parte I

ÉTICA EN LA INTELIGENCIA ARTIFICIAL

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.

2

REVISIÓN DE LA LITERATURA

La literatura sobre ética de la IA se caracteriza por una enorme amplitud y variabilidad de enfoques (Prem, 2023), y una síntesis temática de esta literatura identifica seis conjuntos de problemas no resueltos: la alineación de valores; la equidad y el sesgo; la explicabilidad y la rendición de cuentas; la moral de las máquinas en decisiones de alto riesgo; la gobernanza y la regulación; y la cuestión más profunda de si la ética de la IA es resoluble en absoluto. Una séptima preocupación transversal —la responsabilidad y la rendición de cuentas en el uso cotidiano de la IA— recorre los seis conjuntos, y la revisión cierra examinando tres desafíos críticos a las ambiciones del campo: los límites del aprendizaje continuo, la crítica radical contra la tecnología y los límites epistémicos de la representación y la validez.

2.1 EL PROBLEMA DE LA ALINEACIÓN DE VALORES

El «problema de la alineación de valores» se refiere a garantizar que los sistemas inteligentes persigan únicamente objetivos y actividades beneficiosos para los seres humanos (Sutrop, 2020). El problema enfrenta dos tipos interrelacionados de desafíos, técnicos y normativos: la cuestión técnica pregunta «cómo codificar los valores humanos en la inteligencia artificial», mientras que la cuestión normativa concierne a qué valores deberían guiar la alineación (Sutrop, 2020). Incluso la especificación de los valores objetivo es disputada —la IA podría alinearse plausiblemente con el interés agregado de todos, con prioridad para los peor situados, con una concepción de la virtud del carácter moral o con una ética del reconocimiento mutuo, y ningún candidato individual reúne consenso (Gabriel, 2020). Una segunda dificultad concierne a la legitimidad: dado el abanico de principios

posibles, no está claro quién tiene el derecho de tomar las decisiones de alineación, y sobre qué base (Gabriel, 2020). Subyace a estas cuestiones el hallazgo de que los valores humanos son dinámicos, complejos y difíciles de especificar (Han et al., 2021). Las propuestas técnicas como el aprendizaje por refuerzo inverso y el modelado de recompensas presentan cada una inconvenientes notables, y ningún método único proporciona una solución integral (Singh, 2025). Una revisión sistemática de 172 artículos sobre alineación de valores concluye que la alineación se entiende mejor no como una tarea de especificación única, sino como un proceso continuo entre humanos y agentes autónomos que debe gestionar las exigencias éticas y políticas conflictivas entre grupos (McKinlay et al., 2026).

2.2 EQUIDAD Y SESGO

La toma de decisiones algorítmica está hoy incrustada en la atención médica, el transporte, la educación, las admisiones universitarias, la contratación y los préstamos, y la investigación ha mostrado que tales sistemas pueden ser inherentemente propensos a la injusticia incluso cuando no existe intención de discriminar (Pessach & Shmueli, 2022). Fallos bien documentados ilustran lo que está en juego en la práctica: el chatbot Tay de Microsoft produjo comentarios racistas y sexistas en las 24 horas siguientes a su lanzamiento, un software de predicción de reincidencia asignó puntuaciones de riesgo más altas a afroamericanos que a caucásicos con perfiles similares, y un sistema de recomendación de empleo promovió más oportunidades STEM para hombres que para mujeres (Liu et al., 2022). El campo permanece conceptualmente inestable: los estudios identifican múltiples dimensiones distintas de sesgo —sesgo de datos de entrenamiento, sesgo de modelo, conceptos de equidad en conflicto y ausencia de transparencia de predicción— y ninguna definición única de equidad que satisfaga todos los contextos (Jui & Rivas, 2024). Los mecanismos de mitigación se agrupan convencionalmente en enfoques de preprocesamiento, procesamiento interno y posprocesamiento, aunque identificar qué mecanismo se adecúa a cada escenario sigue siendo un desafío abierto (Pessach & Shmueli, 2022). Los estudios del campo concluyen con desafíos abiertos

articulados como cinco dilemas no resueltos para la investigación sobre equidad (Caton & Haas, 2023), y la dificultad se ve agravada por el hecho de que los modelos profundos cada vez más utilizados en decisiones de alto riesgo vuelven opaco y difícil de auditar su razonamiento (Petersen & Zech, 2024).

2.3 EXPLICABILIDAD, TRANSPARENCIA Y RENDICIÓN DE CUENTAS

La complejidad de las redes neuronales profundas hace que explicar predicciones individuales esté lejos de ser directo —a pesar del conocimiento completo de los pesos y sesgos de una red, la interacción repetida de transformaciones lineales y activaciones no lineales convierte a los modelos en cajas negras (Petersen & Zech, 2024). Las explicaciones son, sin embargo, éticamente necesarias cuando los modelos asignan recursos escasos como trasplantes de órganos, aprobación de créditos, contratación y navegación de vehículos autónomos (Mehta et al., 2023; Petersen & Zech, 2024). Más allá de la explicabilidad se encuentra el problema de la rendición de cuentas. Los estudiosos distinguen una «brecha de rendición de cuentas pública» —la incapacidad de los ciudadanos de obtener explicaciones por las decisiones tomadas por agencias públicas— de una «brecha de rendición de cuentas moral» más amplia, en la que los algoritmos opacos erosionan la capacidad de los agentes humanos de dar sentido a la «lógica» de su propio comportamiento y de explicársela (de Sio & Mecacci, 2021). Esto se conecta con la muy discutida brecha de responsabilidad: situaciones en las que un sistema autónomo actúa sin que nadie —ni el programador, ni el fabricante, ni el operador, ni el propio sistema— sea responsable de la acción y sus consecuencias (Königs, 2022). La brecha es a menudo tan práctica como conceptual: en el comercio de alta frecuencia, la conducción autónoma o la defensa antimisiles automatizada, la toma de decisiones humana es simplemente demasiado lenta para intervenir, lo que hace imposible una «decisión final» humana (Coeckelbergh, 2019). Que la brecha sea genuina, sin embargo, es algo disputado. Algunos sostienen que responsabilizar a una computadora por un daño es un teatro carente de sentido a menos que el sistema posea al menos el mismo nivel de inteligencia,

autoconciencia y sensibilidad moral que un ser humano adulto, dejando un vacío donde debería haber una parte responsable (Goetze, 2022); otros sostienen que la responsabilidad moral es un proceso dinámico y flexible que puede abarcar efectivamente a entidades tecnológicas emergentes sin ninguna brecha basada en la tecnología (Tigard, 2020).

2.4 RESPONSABILIDAD Y RENDICIÓN DE CUENTAS EN EL USO COTIDIANO DE LA IA

El uso cotidiano de la IA —motores de búsqueda, fuentes de redes sociales, herramientas de contratación, calificación crediticia, aplicaciones de salud, navegación— plantea una cuestión práctica de rendición de cuentas: quién ostenta el poder de garantizar el uso correcto, y quién asume la responsabilidad cuando el uso sale mal. La literatura distribuye este poder a lo largo de una cadena de actores en lugar de localizarlo en una única parte (Novelli et al., 2023). Los diseñadores de sistemas son responsables de construir sistemas que satisfagan las necesidades de los usuarios, sean transparentes y explicables, y de proporcionar instrucciones de implementación y divulgar los riesgos potenciales (Liu et al., 2022). Los tomadores de decisiones, que deciden si construir o adoptar un sistema, deben ponderar los beneficios frente a los riesgos y considerar todas las regulaciones pertinentes antes de la implementación; los implementadores de sistemas deben seguir las instrucciones de implementación y capacitar a los usuarios finales; los auditores deben evaluar los sistemas de manera integral y objetiva; y los usuarios finales deben seguir las pautas de uso e informar con prontitud los problemas emergentes (Liu et al., 2022). Las organizaciones internacionales y los reguladores son responsables de guiar correctamente qué puede hacer la IA y dónde residen sus riesgos (Walter, 2024). Esta estructura de roles se corresponde con el ciclo de vida de la IA de diseño, desarrollo e implementación, con agentes identificables individual, corporativa, colectiva o jerárquicamente (Novelli et al., 2023).

En cuanto a la división entre lo «correcto» y lo «incorrecto» en el uso diario, una regla ampliamente compartida es que las empresas y los usuarios asumen juntos la responsabilidad final: las empresas deben hacer todo lo posible por contrarrestar el mal

uso y el abuso de los sistemas de IA, mientras que los usuarios son culpables si crean deliberadamente resultados dañinos o no son suficientemente cuidadosos al automatizar sus tareas (Liu et al., 2022). De manera similar, los desarrolladores son principalmente responsables del diseño y la funcionalidad de una IA —los daños incrustados en el sistema les corresponden a ellos—, mientras que los daños derivados del uso y la implementación corresponden al usuario organizacional (Ryan & Stahl, 2020). Este consenso, sin embargo, no es universal: los profesionales discrepan sobre si los profesionales de la IA asumen la rendición de cuentas por los daños causados por sus sistemas o si esa responsabilidad se comparte con los usuarios (Pant et al., 2024).

Más allá del quién, varios factores determinan si el uso cotidiano resulta correcto o incorrecto. Las decisiones de diseño son de naturaleza social, legal y ética y no pueden dejarse únicamente a ingenieros, fabricantes o usuarios; la IA responsable exige que tales decisiones se informen y permanezcan abiertas a la inspección (Dignum, 2021). La transparencia y la explicabilidad importan porque los sistemas opacos impiden a los usuarios juzgar si una decisión es apropiada, y los mecanismos de rendición de cuentas dependen de hacer la IA inteligible para el público (Radanliev et al., 2024). La calidad de los datos es otro factor: las organizaciones deben ser conscientes de los problemas que implica usar datos deficientes y ser responsabilizadas si se derivan consecuencias dañinas. La educación y el discurso público son igualmente importantes, pues fomentar una cultura del uso moral de la tecnología exige ética en la educación —de lo contrario, los usuarios evalúan la IA sobre la base de sesgos personales más que de una comprensión de cómo funciona (Dignum, 2021). Finalmente, las salvaguardas regulatorias e institucionales —como grupos de ciudadanos, canales de retroalimentación y evaluaciones de impacto— son lo que convierte los principios en práctica efectivamente aplicada.

Una complicación crítica permanece: la autonomía relativa de los sistemas de IA crea una brecha de responsabilidad, lo que hace poco claro quién es responsable cuando ocurre un daño (Königs, 2022). Los estudiosos distinguen una brecha de rendición de cuentas pública —ciudadanos incapaces de obtener explicaciones por decisiones automatizadas de agencias públicas— de

una brecha de rendición de cuentas moral, en la que los algoritmos opacos erosionan la capacidad de los agentes responsables de dar sentido a su propio comportamiento (de Sio & Mecacci, 2021). La misma distribución del poder entre diseñadores, implementadores, auditores y usuarios puede por tanto difuminar la responsabilidad, precisamente donde los daños de la vida diaria son más probables (Lima & Cha, 2020). Si esta brecha es genuina o puede tenderse un puente sobre ella sigue siendo objeto de debate activo (Tigard, 2020). En suma, el poder sobre el uso correcto e incorrecto de la IA se comparte entre diseñadores, implementadores, auditores, usuarios y reguladores —pero el compartir mismo, sin mecanismos claros de atribución, sigue siendo parte del problema no resuelto.

2.5 MORAL DE LAS MÁQUINAS EN DECISIONES DE ALTO RIESGO

Los vehículos autónomos se cuentan entre los primeros robots comerciales con IA incrustada que toman decisiones autónomas, y los accidentes inevitables con consecuencias de vida o muerte no pueden eliminarse por completo —una situación que plantea la cuestión de cómo tales sistemas deberían tomar decisiones éticas, y cómo su comportamiento moral puede hacerse explicable y aceptable para los usuarios (Rhim et al., 2021). El Experimento de la Máquina Moral recopiló aproximadamente 40 millones de respuestas de millones de individuos de 233 jurisdicciones, identificando fuertes preferencias globales por salvar vidas humanas, salvar más vidas y salvar vidas jóvenes (Chen et al., 2022; Lin, 2021). Sin embargo, persiste una profunda variación cultural y demográfica: los encuestados de países de Asia Oriental tendieron a preferir salvar a los ancianos sobre los jóvenes, mientras que los de América del Norte y Europa mostraron la preferencia opuesta, y los estadounidenses expresaron preferencias más fuertes por vehículos autoprotectores cuando viajaban con la familia que los encuestados japoneses (Chen et al., 2022; Lin, 2021). Para los formuladores de políticas, esto plantea un doble problema: experimentos como la Máquina Moral no pueden por sí mismos fundamentar la regulación, pues los factores culturales moldean fuertemente la opinión pública sobre la ética de

los vehículos (Souza & Castañon, 2021), y el marco del dilema hipotético ha sido criticado a su vez por ser excesivamente simplista y determinista, al no representar las condiciones reales de decisión (Kochupillai et al., 2020). Más fundamentalmente, parece no haber una manera fácil de diseñar algoritmos que reconcilien los valores morales y el interés personal a través de culturas con actitudes morales diferentes (Bonnefon et al., 2016), y varios estudiosos sostienen que el marco del problema del tranvía sobrecarga la ética del dilema mientras descuida preocupaciones más realistas como la interacción humano-máquina, la percepción mecánica y la privacidad de los datos (Chen et al., 2022). En respuesta, la literatura se mueve hacia enfoques pluralistas —un pluralismo razonable que exige argumentos de razón pública para las decisiones de diseño (Brändle & Schmidt, 2021), y un pluralismo epistémico moral que trata cada vehículo y el sistema más amplio como capas distintas de consideración moral (Henschke & Arora, 2024).

2.6 GOBERNANZA, REGULACIÓN Y EJECUCIÓN

La gobernanza de la IA se refiere a los marcos, políticas y mecanismos —desde la regulación gubernamental hasta la autogobernanza corporativa— que guían el desarrollo y la operación de los sistemas de IA en consonancia con los principios éticos, las normas sociales y los estándares legales (Walter, 2024). La investigación sobre gobernanza identifica brechas de conocimiento persistentes, incluida la operacionalización insuficiente de los procesos de gobernanza a nivel organizacional y la ausencia de mecanismos de responsabilidad que garanticen el cumplimiento de los principios éticos (Birkstedt et al., 2023). Una agenda de investigación dedicada a la gobernanza técnica de la IA cataloga problemas abiertos: identificar qué dimensiones técnicas se alinean mejor con la regulación (por ejemplo, si los umbrales de cómputo de entrenamiento son medidas de riesgo adecuadas), determinar cómo evaluar las opciones de gobernanza y diseñar mecanismos efectivos de ejecución (Reuel et al., 2024). La regulación misma es un acto de equilibrio: la gobernanza ágil es ampliamente instada porque los marcos regulatorios actuales no dan cuenta adecuadamente del cambio tecnológico rápido

(Dwivedi et al., 2019), pero la regulación excesivamente estricta se contrarresta con advertencias de que puede sofocar la innovación (Dwivedi et al., 2019; Leech et al., 2024). Además, las directrices convergen al nivel de los principios de alto nivel pero divergen marcadamente en la interpretación y la aplicación (Ashok et al., 2021), y un desafío fundamental para cualquier agenda global es equilibrar la armonización transnacional con el respeto por el pluralismo cultural y moral (Jobin et al., 2019).

2.7 LA METACUESTIÓN: ¿PUEDE RESOLVERSE LA ÉTICA DE LA IA?

Un último conjunto concierne a la viabilidad de toda la empresa. Se han propuesto docenas —de hecho, más de un centenar— de marcos éticos, modelos de proceso y herramientas, y sin embargo muchos son demasiado abstractos para traducirse fácilmente en diseños concretos de sistemas (Prem, 2023). El «principlismo» resultante se centra generalmente en un pequeño conjunto de principios como la beneficencia, la no maleficencia, el respeto por la autonomía y la justicia, reflejando la ética biomédica, pero los principios por sí solos no especifican cómo deberían aplicarse en un contexto particular (Jedličková, 2024; Prem, 2023). Los esfuerzos por operacionalizarlos son extensos —una revisión sistemática identifica 106 herramientas, otra 39—, y sin embargo la mera plétora de principios resulta abrumadora y confusa para los desarrolladores, y la mayoría del personal técnico los desconoce (Ortega-Bolaños et al., 2024). Los comentaristas observan que la ética de la IA ha evolucionado a través de fases, de los principios de alto nivel (2016–2019) a una fase técnica de equidad-rendición de cuentas-transparencia y ahora una fase marcada por el movimiento «del qué al cómo», centrada en mecanismos de gobernanza, auditoría y estándares (Ayling & Chapman, 2021). Los críticos advierten que las pautas de marcar casillas no deben ser los únicos «instrumentos» de la ética de la IA, y piden una transición hacia enfoques sensibles a la situación y basados en la virtud (Hagendorff, 2020). Incluso el objetivo último es disputado: el florecimiento humano se propone como el principio ético orientador, y sin embargo no es trivial determinar qué constituye el florecimiento ni cómo contribuye la tecnología a él (Stahl et al.,

2020), y se ha acusado a la ética de la IA de promover el bienestar de algunos mientras socava activamente el de otros (Munn, 2022).

2.8 EL DILEMA ESTABILIDAD-PLASTICIDAD: APRENDIZAJE CONTINUO Y GENERALIDAD DE LA IA

2.8.1 *Lo que el artículo realmente establece*

El dilema estabilidad-plasticidad se define clásicamente como la necesidad de que los modelos sean «suficientemente estables para retener el conocimiento aprendido de clases vistas previamente, y suficientemente plásticos para aprender conceptos de clases nuevas», mientras que el ajuste fino ingenuo sufre olvido catastrófico porque la alta plasticidad sobrescribe parámetros importantes para las tareas antiguas (Kim & Han, 2023). El estudio fundacional de este dilema, sin embargo, es diagnóstico más que prescriptivo, y su afirmación central es casi lo opuesto a una solución: la mayoría de los algoritmos de aprendizaje incremental por clases «favorecen fuertemente la estabilidad sobre la plasticidad», hasta el punto de que el extractor de características entrenado sobre las clases iniciales no es «menos eficaz que el del modelo incremental final» (Kim & Han, 2023). Midiendo la plasticidad al reentrenar clasificadores en cada etapa y la estabilidad mediante la Alineación de Núcleos Centrada, los autores constatan que la mayoría de los modelos «acumulan poco conocimiento nuevo en sus representaciones de características a lo largo de las etapas incrementales» —vencen el olvido catastrófico simplemente negándose a aprender (Kim & Han, 2023). La demostración más tajante es su «Exploit» deliberadamente degenerado: congela por completo el extractor de características y entrena solo una pequeña cabeza, y sin embargo alcanza un 67,2 % de precisión incremental promedio y un 60,5 % de precisión final en el escenario de 500 clases —competitivo con todos los métodos salvo DER y pDER, frente a un techo «Oracle» de 72,8 %/70,8 % (Kim & Han, 2023). El veredicto de los autores: las métricas CIL estándar «ofrecen una falsa ilusión de progreso» (Kim & Han, 2023). Solo los métodos de aislamiento de parámetros —DER, que apila un nuevo extractor de características por etapa— exhiben plasticidad genuina, mostrando una precisión creciente por

etapa donde otros cinco algoritmos permanecen planos y solo el Oracle crece casi linealmente (Kim & Han, 2023). Pero DER exige N+1 pasadas hacia adelante para N pasos, y la modificación del artículo (pDER) solo mitiga ese costo (Kim & Han, 2023). La cuestión teórica central —si un equilibrio fuerte entre estabilidad y plasticidad es siquiera posible sin añadir parámetros— permanece abierta, aunque el teorema de Raghavan y colaboradores garantiza que puede existir un punto de equilibrio estable entre el olvido y la generalización (Kim & Han, 2023).

2.8.2 *La conexión humana: ¿es el dilema «epistemológicamente sólido»?*

El dilema es empíricamente real en los seres humanos: ha sido «ampliamente estudiado tanto para sistemas computacionales como biológicos» (Parisi et al., 2018), y la teoría de los sistemas de aprendizaje complementarios explica cómo la biología lo resuelve —un sistema hipocampal rápido que codifica rápidamente los detalles, emparejado con un sistema neocortical lento que extrae la estructura mediante aprendizaje entrelazado y reproducido (Kumaran & McClelland, 2012; McClelland et al., 1995). Los humanos sí muestran fenómenos de interferencia (por ejemplo, aprender asociaciones A–C tras haber codificado A–B), aunque «mucho menos dramáticos» que las redes (Saxe et al., 2020). Pero el dilema es una restricción arquitectónico-cognitiva, no un principio epistémico. La epistemología concierne a la verdad, la justificación y la creencia; el dilema concierne a la retención frente a la adaptación de la información almacenada. Lo que la teoría sí apoya es la afirmación más débil: la disyuntiva no es un artefacto del aprendizaje profundo sino una condición general de los sistemas de aprendizaje finitos con fundamento matemático (Kim & Han, 2023) —de modo que la analogía humana se sostiene en el sentido de ser principiada e inevitable, no en el sentido epistemológico.

2.8.3 *¿Autoriza esto la inteligencia artificial general — y vuelve irrelevante el enfoque lógico?*

No, en ninguno de los dos casos, por tres razones independientes. Primero, los resultados positivos son estrechos y en parte ilusorios. El escenario es la clasificación de imágenes incremental por clases; incluso el techo del Oracle se sitúa en ~71–73 % de precisión (Kim & Han, 2023). El trabajo más amplio coincide en que el aprendizaje incremental por clases con entradas complejas «sigue siendo un problema abierto» a menos que los datos se almacenen explícitamente (van de Ven et al., 2020), que «sigue existiendo una enorme brecha entre los avances existentes y las aplicaciones realistas» (Wang et al., 2023), y que la plasticidad misma se degrada con el entrenamiento prolongado —el problema de la «pérdida de plasticidad» (Dohare et al., 2024). La percepción continua está muy lejos de la inteligencia general, que exige razonamiento, lenguaje, planificación, explicación y juicio de valor —competencias que esta línea de trabajo no toca. Segundo, las soluciones genuinas son computacionalmente explosivas, y las honestas son trampas. El único enfoque que muestra estabilidad y plasticidad reales multiplica el cómputo por el número de etapas (Kim & Han, 2023), y el propio «Exploit» del artículo muestra con qué facilidad se manipulan los puntos de referencia (Kim & Han, 2023). Una IAG construida sobre esta base heredaría un problema de escalabilidad no resuelto y una cuestión teórica no zanjada (Kim & Han, 2023). Tercero, la biología es evidencia a favor de la hibridación, no en contra de la lógica. La solución del propio cerebro es una división del trabajo: un sistema rápido basado en episodios y un sistema lento extractor de estructura (Kumaran & McClelland, 2012; McClelland et al., 1995). El enfoque lógico aborda una competencia completamente distinta —la representación del conocimiento como simbolización formal de proposiciones creídas y el razonamiento como su manipulación formal (Lee, 2024), con los límites de expresividad-tratabilidad tratados antes en esta revisión, y con la maquinaria formal-semiótica del hexágono (Servín, s.f.-b) ocupándose del estatus epistémico y normativo de las afirmaciones. El dilema estabilidad-plasticidad trata del mantenimiento de la memoria; la lógica trata de la implicación y la explicación. Ninguno vuelve

irrelevante al otro —y el cerebro humano, que enfrenta el dilema y aun así razona lógicamente, es el contraejemplo permanente. La conclusión defendible es que el dilema es una restricción genuina y compartida del aprendizaje biológico y artificial —la analogía humana se sostiene empíricamente—, pero esto no autoriza ni la IAG ni la irrelevancia de la lógica; a lo sumo refuerza el caso a favor de las arquitecturas híbridas, que es, no incidentalmente, lo que el cerebro ya es.

2.9 CRÍTICA RADICAL CONTRA LA TECNOLOGÍA: EL DESAFÍO DE KACZYNSKI

El manifiesto de Theodore Kaczynski —*La sociedad industrial y su futuro*— ha sido ahora objeto de un cuidadoso análisis filosófico e histórico, incluso mientras sus métodos como autor son universalmente repudiados (Fleming, 2021; Moen, 2018). Aquí se examina como *discurso*, no como interlocutor filosófico en pie de igualdad. Barnett muestra que el manifiesto combina apelaciones emocionales con «evidencia verificable» (vigilancia, concentración mediática, riesgos ambientales) dentro de una dicotomía tecnología/naturaleza simplificada a propósito, y que la violencia fue para su autor una estrategia de difusión (Barnett, s.f., pp. 62-65). La lección metodológica es que la fuerza persuasiva de un texto es independiente de la verdad de sus afirmaciones: reconocer algunos de sus hechos no obliga a aceptar sus inferencias ni, menos aún, su programa. Con esa cautela, su desafío al proyecto reformista de la ética de la IA merece examen, porque formula de manera extrema una pregunta legítima: si la reforma desde dentro puede fracasar aun cuando todos sepan que es necesaria.

2.9.1 *Lo que exige el manifiesto*

Tres afirmaciones estructuran *La sociedad industrial y su futuro*:

1. La tecnología destruye el «proceso de poder». Los seres humanos necesitan metas y la libertad de perseguirlas, enfrentando los riesgos de sus propias decisiones; la transferencia progresiva de la toma de decisiones a sistemas

impersonales y deshumanizadores impide a los individuos realizar y perseguir estas necesidades fundamentales (Cappuccio et al., 2023). Se opuso a la tecnología principalmente por razones evolutivo-psicológicas —la sociedad moderna obliga a las personas a vivir bajo condiciones radicalmente diferentes de aquellas bajo las cuales evolucionó la raza humana (Fleming, 2024).

2. Las máquinas terminarán por tener el control. Su advertencia más conocida: «la raza humana podría fácilmente permitirse derivar hacia una posición de tal dependencia de las máquinas que no tendría más opción práctica que aceptar todas las decisiones de las máquinas... en esa etapa las máquinas estarán en control efectivo» (Yampolskiy, 2022).

3. La reforma es imposible; solo la disolución basta. La industrialización ha sido terrible para la humanidad, y solo un retorno al modo de vida preindustrial puede remediarlo —por cualquier medio necesario, en su opinión (Moen, 2018).

2.9.2 *Dónde el documento coincide parcialmente*

La brecha de responsabilidad es la coincidencia más estrecha. En la sección «Responsabilidad y rendición de cuentas en el uso cotidiano de la IA», el documento señala que la autonomía relativa de la IA crea una brecha de responsabilidad y que distribuir el poder entre diseñadores, implementadores, auditores y usuarios «puede también difuminar la responsabilidad, precisamente donde los daños de la vida diaria son más probables». El diagnóstico se aproxima a la segunda afirmación de Kaczynski —que las decisiones migran a las máquinas más rápido de lo que los humanos pueden mantener un control inteligible sobre ellas (Yampolskiy, 2022)—, aunque la coincidencia se limita a la descripción del problema y no se extiende a sus causas ni a sus remedios. Ambos textos diagnostican una erosión de la agencia humana efectiva —el manifiesto atribuye a ello un malestar psicológico generalizado (Fleming, 2024), mientras que el documento lo trata como un déficit de rendición de cuentas.

La moral de las máquinas en decisiones de alto riesgo. La crítica del documento a la ética del problema del tranvía —que el marco del dilema sobrecarga los escenarios artificiales y no representa las condiciones reales de decisión— converge con la objeción más profunda del manifiesto, que se dirige a los procedimientos formales e impersonales de decisión que reemplazan por completo el juicio humano (Cappuccio et al., 2023). Kaczynski leería el propio proyecto de «programar la ética en las máquinas» como un error categorial: la agencia ética exige enfrentar las consecuencias de las propias decisiones, que es exactamente lo que la delegación a las máquinas erosiona.

La metacuestión (¿puede resolverse la ética de la IA?). El documento documenta el fracaso del principlismo, la proliferación de herramientas inconmensurables y la noción disputada del florecimiento humano. La respuesta de Kaczynski a esa frustración es la radical: no puede —no dentro del sistema, porque la complejidad y el impulso del sistema desbordan cualquier correctivo. Moen constata que «aunque su teoría en su conjunto debería rechazarse», algunas preocupaciones puntuales sobre el desarrollo tecnológico «han llegado recientemente a ser compartidas por destacados defensores de la mejora humana», incluidos Bostrom y Savulescu (Moen, 2018). La observación es de Moen y está acotada: la convergencia se limita a la preocupación por la deriva inducida por la IA (Yampolskiy & Fox, 2012), no a sus fundamentos. La investigación en seguridad atribuye los riesgos a la mala especificación de objetivos y a la instanciación perversa de metas formalmente correctas, no a un supuesto impulso de dominación de los artefactos ni a un determinismo evolutivo-psicológico.

2.9.3 *Dónde el documento diverge fundamentalmente*

La equidad y el sesgo constituyen el desacuerdo más agudo. El documento trata el sesgo algorítmico como un defecto que debe corregirse mediante mejores datos, mejores métricas y mejor mitigación. Para Kaczynski, las normas igualitarias, imparciales y no violentas que respaldan tal corrección son ellas mismas síntomas de «sobresocialización» —normas inculcadas precisamente para que los individuos funcionen sin fricción dentro de sistemas complejos de producción, e imponiendo una carga que «produce una

sensación de constricción e impotencia» (Fleming, 2024). Desde su perspectiva —que este documento no comparte—, la ingeniería de la equidad le parecería el sistema reparando su propia legitimidad en lugar de abordar la totalidad deshumanizadora (Galliott, 2017).

Gobernanza, regulación y ejecución. Toda la sección de gobernanza del documento presupone lo que Kaczynski niega de plano: que el sistema puede reformarse desde dentro. Sus hallazgos —las directrices divergen en la interpretación, la operacionalización fracasa, la ejecución está ausente— son precisamente la evidencia que Kaczynski citaría para su conclusión estructural de que la corrección fragmentaria es fútil (Moen, 2018). El documento ve una brecha de gobernanza; él ve una ilusión de gobernanza.

Alineación de valores. El documento pregunta qué valores y qué autoridad deberían guiar la alineación (Gabriel, 2020; Singh, 2025). Kaczynski respondería que la pregunta está amañada: el sistema industrial-tecnológico no puede ser puesto al servicio de los valores humanos porque subordina los fines humanos a sus propias exigencias funcionales. Un análisis filosófico de su posición concluye en efecto que los usuarios de sistemas tecnológicos complejos se enfrentan a un trilema —«tecnología encarcelada controlada por quienes se hallan dentro de las estructuras de poder opresivas existentes», dedicar la propia vida a dominar los sistemas, o «rebelarse contra dicho sistema» (Galliott, 2017). La alineación como ingeniería no figura en el menú de Kaczynski.

2.9.4 *El veredicto fundamental*

La coincidencia es diagnóstica, no prescriptiva. Donde el documento y el manifiesto se encuentran es en el diagnóstico: la complejidad desborda la comprensión humana, la autonomía difumina la responsabilidad, la agencia se erosiona y los principios no logran constreñir la práctica. Donde se separan irreconciliablemente es en la prescripción: el documento busca gobernar, explicar, alinear y regular la tecnología desde dentro; Kaczynski exigía su abolición. Incluso su única premisa de apariencia compartida es discutible —los críticos señalan que su suposición de que las máquinas «buscarán dominar» atribuye a los artefactos

un rasgo humano y evolucionado (el deseo de dominación) que estos no poseen (Yampolskiy & Fox, 2012)—, y su teoría en su conjunto ha sido juzgada filosóficamente implausible, siendo su violencia el producto de una fe ciega en una utopía antitecnológica (Lagarde, 2019; Moen, 2018). Su crítica funciona, no obstante, como una condición de frontera, y conviene formularla de modo que sea evaluable y no fatalista. No afirma que el sistema no pueda reformarse, sino que traslada la carga de la prueba: la ética de la IA debe mostrar que la gobernanza y la rendición de cuentas funcionan en la práctica. Esa exigencia tiene una versión analítica que no requiere ni determinismo ni utopía antitecnológica. Falk y Tsoukalas demuestran que, cuando cada empresa captura todo el ahorro de automatizar pero soporta solo una fracción de la pérdida de demanda resultante, la automatización excesiva es una estrategia dominante: *conocer el riesgo no basta para corregirlo*, porque los incentivos tienen la estructura de un dilema del prisionero; pero el mismo análisis identifica instrumentos —un impuesto pigouviano a la automatización— que restauran el óptimo cooperativo (Falk & Tsoukalas, 2026, Proposiciones 1 y 5). La pregunta de Kaczynski se transforma así en una pregunta empírica e institucional: qué evidencia mostraría que la gobernanza funciona —principios operacionalizados y auditados, sanciones efectivamente aplicadas, vías de impugnación usadas por los afectados, supervisión humana que no se erosiona a medida que los sistemas parecen más fiables— y qué estructuras de incentivos la hacen posible. Si esa evidencia existe, la tesis abolicionista queda refutada en ese punto; si no existe, la reforma tiene la carga de la prueba. Su argumento del «proceso de poder» sigue siendo, por otra parte, el concepto más cercano a las secciones de este documento sobre la agencia y la autonomía humanas, y la influencia del manifiesto en el radicalismo antitecnológico contemporáneo se ha rastreado hasta una síntesis de Jacques Ellul, Desmond Morris y Martin Seligman (Fleming, 2021).

2.10 LOS LÍMITES EPISTÉMICOS DE LA REPRESENTACIÓN: POR QUÉ LAS AFIRMACIONES DE «CUALQUIER PROBLEMA» SON INJUSTIFICADAS

Un desafío adicional a las ambiciones de los sistemas de IA, epistémico más que político, se sigue de la propia teoría de la representación del conocimiento. El análisis de Brachman (1979) dividió el nivel epistemológico de la representación en cuatro capas distintas: la *implementacional*, relativa a las estructuras de datos; la *lógica*, relativa a la lógica simbólica, las proposiciones, los predicados, las variables y los cuantificadores; la *epistemológica*, relativa a los tipos de concepto, subtipos, herencia y relaciones estructurantes; y la *conceptual*, relativa a las relaciones semánticas, los roles lingüísticos, los objetos y las acciones (Sowa, 1999). El nivel epistemológico es donde la realidad se talla en categorías —y cualquier tallado de este tipo es una elección más que un dato (Sowa, 1999). Según la definición canónica, la representación del conocimiento es el uso de símbolos formales para representar una colección de proposiciones creídas por algún agente putativo, de las cuales solo un número finito se representa alguna vez, y el razonamiento es la manipulación formal de esos símbolos para producir representaciones de nuevas proposiciones (Lee, 2024). El razonamiento, dicho de otro modo, opera sobre un modelo, no sobre el mundo.

La representación es, por tanto, inherentemente parcial y relativa a propósitos. La misma sustancia puede representarse con axiomas diferentes, incluso contradictorios, en distintos contextos y para distintos fines: el ingeniero trata la lechada de carbón como continua para el transporte y como grumosa para la separación, y ambas representaciones son legítimas (Sowa, 1999). La «precisión» de un sistema es, en consecuencia, precisión bajo una representación elegida —relativa a los axiomas, las categorías y la granularidad que los diseñadores humanos seleccionaron. Afirmar la precisión sin calificaciones comete precisamente la confusión categorial que los niveles de Brachman fueron diseñados para exponer: confunde el modelo con el mundo (Sowa, 1999).

El razonador implementado ha pagado además ya un precio de expresividad-tratabilidad. La literatura es explícita en que

incluso pequeñas dosis de poder expresivo pueden acarrear un costo computacional significativo, y que las cuestiones de qué controla la dificultad computacional, hasta dónde puede empujarse la expresividad y cuándo bastan lenguajes inexpresivos pero tratables siguen en gran medida sin respuesta (Lee, 2024). La lógica de primer orden es indecidible —demostrado independientemente por Church y Turing en 1936—, y los investigadores han tenido que retirarse a lenguajes deliberadamente más débiles porque la búsqueda de implicaciones es computacionalmente desesperanzada en general (Lee, 2024). El propio análisis de nivel de conocimiento de Newell pone en primer plano la complejidad computacional intrínseca de la relación de implicación como una cuestión de adecuación central para cualquier lenguaje de representación (Sowa, 1999). Un sistema capaz de «cualquier problema» exigiría así capacidades que el campo no sabe cómo construir, en principio más que meramente en la práctica.

Dos límites adicionales cierran las vías restantes hacia la afirmación fuerte. Primero, la equidad no es una propiedad epistémica: los niveles de Brachman describen cómo se estructura el conocimiento, no cómo se fundamentan los valores, y el nivel epistemológico codifica tipos de concepto y herencia sin vocabulario para la justicia, el daño o el mérito (Sowa, 1999). La selección de qué representar, con qué granularidad y bajo qué axiomas es un acto humano cargado de valor —pero la maquinaria que razona sobre esa selección es silenciosa respecto de los valores. Afirmar que un sistema resuelve problemas «con equidad» es afirmar que una propiedad normativa se sigue de la manipulación formal de símbolos, una inferencia que la propia distinción conocimiento/símbolo bloquea. Segundo, la resolución de problemas humana se apoya en la intuición, que concierne a los particulares más que a las verdades generales, requiriendo estas últimas inferencia inductiva o abductiva (Servín, s.f.-a); el razonamiento formal sobre una base de conocimiento es deductivo, y la brecha entre lo representado y lo creído no es cerrada por el marco representacional. Ni existe lógica privilegiada alguna que la cierre: la lógica interna del hexágono lógico-semiótico no es necesariamente booleana sino dependiente del modelo descrito por sus categorías (Servín, s.f.-b). No hay lógica canónica, ni granularidad canónica, ni modelo canónico —de ahí que no

haya ruta desde ninguna base de conocimiento hasta «cualquier problema, con precisión».

La cuestión ética se sigue directamente. Las afirmaciones injustificadas de precisión y equidad inducen a error a usuarios, reguladores y público sobre la confiabilidad de los sistemas; transfieren la confianza del juicio humano a los modelos; y diluyen la rendición de cuentas precisamente donde los daños son más probables, pues el marco de «el sistema lo resolvió» oscurece las decisiones humanas —qué axiomas, qué categorías, qué granularidad, qué valores— que en realidad determinaron el resultado (de Sio & Mecacci, 2021; Königs, 2022). Este es el puente epistémico hacia el análisis de la brecha de responsabilidad desarrollado antes: si la competencia de un sistema es siempre competencia-relativa-a-un-modelo, entonces quien eligió el modelo asume una responsabilidad que la retórica de la precisión autónoma borra. La afirmación defendible no es que la IA pueda resolver con precisión y equidad cualquier problema que los humanos puedan resolver, sino la afirmación más débil y fiel a la representación: la IA puede generar implicaciones de una representación elegida, con una granularidad elegida, bajo una lógica elegida, dentro de límites de tratabilidad —y cada una de esas elecciones es una decisión humana cargada de valor que el propio sistema no puede tomar (Lee, 2024; Sowa, 1999).

2.11 VALIDEZ, INTERPRETACIÓN Y LA ÉTICA DE LA AUTORIDAD EPISTÉMICA DELEGADA

El análisis epistemológico de la representación plantea una cuestión ulterior de legitimidad científica y ética: si los resultados de los sistemas de IA pueden tratarse propiamente como interpretaciones válidas de los datos que procesan. La distinción lógica entre validez y solidez proporciona el vocabulario preciso. Un argumento deductivo es válido si, fueran verdaderas todas sus premisas, la conclusión tendría que ser verdadera; es sólido solo si es válido *y* todas sus premisas son de hecho verdaderas (Servín, s.f.-a). Aplicada a los sistemas de IA, la distinción es decisiva. Un sistema de razonamiento bien construido deriva sus resultados correctamente de sus entradas representadas —manipulación de símbolos, inferencia o generalización estadística

a partir de los datos que se le dieron—, y en ese sentido restringido un motor de IA puede ser un dispositivo *válido*. La solidez, sin embargo, no es un producto entregable de la maquinaria: exige que las premisas sean verdaderas, y las premisas de un sistema de IA son su representación, a saber, las categorías, los axiomas, la granularidad, los datos y las variables objetivo seleccionados por diseñadores humanos (Sowa, 1999). Puesto que la representación del conocimiento codifica solo un número finito de las proposiciones creídas por algún agente, y el razonamiento es la manipulación formal de esos símbolos, el sistema opera sobre un modelo más que sobre el mundo (Lee, 2024). La verdad sobre el mundo nunca es derivada por el sistema; es importada a través de la representación. Un sistema cuyas premisas son falsas —datos sesgados, una variable objetivo mal especificada, una granularidad inapropiada— es formalmente válido y materialmente insólido.

El estatus epistémico de la producción mecánica se relativiza además por el marco interpretativo en el que se coloca. Lo que cuenta como justificación depende de la epistemología que uno adopte: si hay epistemologías distintas, el término «necesidad epistémica» se vuelve relativo, mientras que la necesidad lógica debe entenderse según las leyes de alguna lógica —y qué lógica es, es por sí mismo una cuestión abierta (Servín, s.f.-a). Ninguna lógica canónica existe para fijar el estatus de los resultados: la lógica interna del hexágono lógico-semiótico no es necesariamente booleana sino dependiente del modelo descrito por sus categorías (Servín, s.f.-b). La interpretación no es, por tanto, suministrada por el sistema. En teoría de modelos, la herramienta principal de la semántica es el concepto de valor de verdad, y sin embargo la referencia —especialmente en el lenguaje natural— es uno de los temas más problemáticos de la semántica (Servín, s.f.-b). Quién fija la referencia de los símbolos del sistema, y contra qué modelo, sigue siendo una decisión humana externa a la maquinaria.

La distinción entre validez y solidez es, con todo, deductiva, y su aplicación a los sistemas de aprendizaje estadístico es una analogía que debe precisarse. Un razonador simbólico deriva sus conclusiones de premisas; una red entrenada no deduce sus salidas, sino que generaliza por semejanza a partir de ejemplos. Para estos sistemas el marco adecuado es el de la *abducción* y la

intuición artificial descrita antes: la salida es una hipótesis plausible que solo se convierte en conocimiento cuando se reconstruye como argumento y se contrasta. Su riesgo característico no es únicamente la premisa falsa (la falta de solidez), sino el *acierto accidental*: un sistema puede producir una respuesta verdadera apoyándose en un rasgo espurio de los datos, y quien la acepta por la fiabilidad general del sistema tiene una creencia verdadera y en apariencia justificada que no es conocimiento —un caso de Gettier (Servín, s.f.-a). De ahí se siguen grados que no deben confundirse: una *salida* se convierte en *evidencia* cuando se conocen su procedencia y el tipo de apoyo que ofrece; la evidencia, en *aceptación justificada* cuando un sujeto responsable la integra en un argumento evaluable; la aceptación, en *conocimiento* cuando es verdadera, no accidental y está corroborada intersubjetivamente; y el conocimiento solo apoya una *decisión* cuando se añaden premisas normativas explícitas y la consideración de los afectados. La corrección de un procedimiento, además, solo puede medirse contra una semántica independiente de él, no contra «lo que el código hace» (Brachman & Levesque, 1984).

La afirmación de que las interpretaciones generadas por la IA son válidas simplemente porque la IA se apoya en métodos científicos comete por tanto un error categorial: sustituye la validez técnica por la validez interpretativa y normativa. Un trabajo reciente distingue tres dimensiones a lo largo de las cuales debe juzgarse la validez de los resultados de la IA —la corrección técnica, la validez interpretativa (si los resultados pueden entenderse y aplicarse de manera significativa) y la validez normativa (la equidad y el respeto por los estándares éticos) (Hakim et al., 2026). La «validez científica» se refiere característicamente solo a la dimensión técnica, y el rigor técnico no purifica el resultado. La literatura sobre IA responsable hace el mismo punto en negativo: las conclusiones de la IA descansan sobre estadística inferencial y aprendizaje automático que producen conocimiento probabilístico e incierto y son, por tanto, falibles; la evidencia puede ser inconcluyente, y los sistemas de caja negra inescrutables rompen la conexión entre los datos usados, el modo en que se usaron y la conclusión generada (Trocin et al., 2021). Cuando la derivación es inescrutable, la solidez no puede verificarse sino solo suponerse. Incluso la «precisión» —el valor que la inferencia

presume neutral y absoluto— conlleva un componente axiológico y pragmático constitutivo, de modo que la cuestión operativa es siempre la precisión *para quién*, y para qué propósito (Russo et al., 2023).

La dimensión ética se sigue directamente. Afirmar que toda interpretación es válida porque el método es científico convierte un resultado demostrable y relativo al modelo en una certificación dirigida al mundo, induciendo a error a usuarios, reguladores y público sobre la confiabilidad de los sistemas. El sesgo y el error en los sistemas de IA no meramente distorsionan resultados individuales; socavan la validez y la confiabilidad del conocimiento científico mismo, con consecuencias potencialmente desastrosas para la salud pública, la seguridad y la política social —y la ética de la investigación científica exige que tales errores se identifiquen, describan y corrijan en lugar de disculparlos por el rigor del método (Resnik & Hosseini, 2024). La misma retórica borra las decisiones humanas —qué datos, qué categorías, qué granularidad, qué valores— que en realidad determinaron el resultado, y disuelve así la asignación rastreable de responsabilidad a lo largo del ciclo de vida que la rendición de cuentas exige (Novelli et al., 2023). Cuando «el sistema lo resolvió correctamente» se afirma sin calificaciones, los diseñadores, implementadores y usuarios que eligieron las premisas quedan relevados de la rendición de cuentas, reproduciendo la brecha de responsabilidad en la que un sistema autónomo actúa sin que nadie sea responsable de la acción y sus consecuencias (de Sio & Mecacci, 2021; Königs, 2022). La alternativa defendible es tratar la validez no como una propiedad que un sistema posee de una vez y para siempre, sino como un juicio relativo a propósitos, concedido solo para los usos previstos y mantenido abierto a revisión —una práctica responsable y reflexiva más que un certificado de autoridad científica.

3

METODOLOGÍA

Esta revisión sigue una síntesis narrativa y temática de la literatura revisada por pares sobre ética de la IA. Las fuentes se identificaron mediante búsquedas estructuradas en bases de datos académicas que cubren revistas, actas de congresos y preprints, complementadas por búsquedas dirigidas en la biblioteca personal de materiales de investigación del autor. Los trabajos fueron cribados según su relevancia directa para los desafíos éticos planteados por la inteligencia artificial, con prioridad para las revisiones sistemáticas, los artículos de posición de alto impacto y los estudios empíricamente fundamentados. La literatura incluida fue codificada temáticamente, y los códigos resultantes se organizaron en los conjuntos de problemas presentados en la revisión de la literatura, junto con las secciones críticas y epistemológicas que acotan las ambiciones del campo. Puesto que la revisión es interpretativa más que estadística, no se realizó ninguna síntesis cuantitativa o meta-analítica; los hallazgos se reportan cualitativamente. La limitación principal de este enfoque es su dependencia de la literatura secundaria y del juicio del revisor al agrupar debates heterogéneos bajo temas compartidos.

4

RESULTADOS

La síntesis arroja seis hallazgos principales. Primero, la alineación de valores sigue sin resolverse tanto en el frente técnico como en el normativo: no hay un método acordado para codificar los valores humanos, ni un conjunto acordado de valores que codificar, ni una autoridad acordada que elegir entre ellos (Gabriel, 2020; Sutrop, 2020). Segundo, la investigación sobre equidad carece de una definición asentada de equidad, enfrenta múltiples fuentes de sesgo irreductibles y aún no ha producido una mitigación de extremo a extremo que abarque datos, modelo e implementación (Jui & Rivas, 2024; Pessach & Shmueli, 2022). Tercero, la explicabilidad y la rendición de cuentas están ligadas: los modelos opacos socavan tanto la justificación como la responsabilidad, y la existencia y naturaleza de la brecha de responsabilidad siguen siendo filosóficamente disputadas (Königs, 2022; Petersen & Zech, 2024; Tigard, 2020). Cuarto, el diseño de la moral de las máquinas está bloqueado por obstáculos empíricos y normativos a la vez —la heterogeneidad cultural de las preferencias morales y la ausencia de cualquier marco sobre el que el público y los reguladores puedan ponerse de acuerdo (Bonnefon et al., 2016; Lin, 2021). Quinto, la gobernanza va a la zaga del progreso tecnológico, con problemas abiertos en la operacionalización, la auditoría, el cumplimiento y la identificación de dimensiones regulatorias objetivo (Birkstedt et al., 2023; Reuel et al., 2024). Sexto, la viabilidad misma de una solución está en cuestión, dada la abstracción de los principios, la proliferación de herramientas sin conmensurabilidad y el fundamento normativo no resuelto de la propia empresa (Prem, 2023; Stahl et al., 2020).

Un séptimo hallazgo transversal concierne al uso cotidiano: la responsabilidad se distribuye entre diseñadores, tomadores de

decisiones, implementadores, auditores, usuarios, organizaciones, reguladores y organismos internacionales a lo largo del ciclo de vida de diseño–desarrollo–implementación (Liu et al., 2022; Novelli et al., 2023; Ryan & Stahl, 2020; Walter, 2024), pero el compartir mismo puede difuminar la responsabilidad, y si existe una brecha de responsabilidad genuina sigue siendo disputado (de Sio & Mecacci, 2021; Königs, 2022; Tigard, 2020). Finalmente, tres desafíos críticos acotan las ambiciones más fuertes del campo. La literatura sobre aprendizaje continuo muestra que el progreso aparente en el dilema estabilidad-plasticidad refleja a menudo una estabilidad comprada a expensas de la plasticidad (Kim & Han, 2023). La crítica radical contra la tecnología niega que el sistema tecnológico pueda reformarse desde dentro (Moen, 2018; Yampolskiy, 2022). Y la epistemología de la representación del conocimiento muestra que las afirmaciones de precisión y equidad universales son injustificadas: un sistema genera implicaciones de representaciones elegidas por humanos, y la solidez no puede ser certificada por ninguna maquinaria que razone sobre ellas (Lee, 2024; Servín, s.f.-a; Sowa, 1999).

5

DISCUSIÓN

De estos hallazgos emergen dos tensiones transversales. La primera es la brecha principio–práctica: incluso donde las partes interesadas se ponen de acuerdo sobre principios de alto nivel, la traducción a mecanismos accionables de diseño, gobernanza y ejecución sigue siendo incompleta (Ashok et al., 2021; Morley et al., 2019). La segunda es la tensión entre armonización y pluralismo: el acuerdo global sobre una IA ética es deseable, pero los intentos de imponer estándares universales corren el riesgo de obliterar el pluralismo cultural y moral (Jobin et al., 2019), y las acusaciones de hegemonía cultural en la ética de la IA siguen vivas (Goffi, 2021). Los hallazgos implican también que los problemas más difíciles del campo no son puramente técnicos. Los debates sobre alineación, equidad y rendición de cuentas están todos subdeterminados por la ingeniería sola, lo que sugiere que el progreso depende de la colaboración interdisciplinaria entre tecnólogos, éticos, formuladores de políticas y el público (Hagendorff, 2020). Para la investigación, la agenda apunta a desarrollar métodos éticos sensibles a la situación en lugar de listas de verificación universales (Hagendorff, 2020); para la práctica, a incrustar la ética a lo largo de todo el ciclo de vida de la IA (Prem, 2023); y para la política, a mecanismos regulatorios adaptativos que puedan seguir el ritmo del cambio rápido sin sofocar la innovación (Dwivedi et al., 2019).

Las perspectivas críticas adicionales agudizan estas tensiones. El desafío de Kaczynski, leído como discurso y no como programa (Barnett, s.f.), funciona como una condición de frontera evaluable: la gobernanza, la alineación y la rendición de cuentas deben demostrarse en la práctica, y cuando los incentivos tienen la estructura de un dilema colectivo, saber que un riesgo existe

no basta para corregirlo (Falk & Tsoukalas, 2026; Moen, 2018). La respuesta no es la abolición, sino el diseño de instituciones e incentivos cuya eficacia pueda verificarse. La crítica epistemológica, a la inversa, constriñe al campo desde dentro: los sistemas generan implicaciones de representaciones elegidas, y cada elección —de lógica, de granularidad, de categoría, de valor— es una decisión humana que ningún progreso técnico puede convertir en un oráculo (Lee, 2024; Sowa, 1999). Juntas, las dos críticas niegan al campo sus ambiciones más fuertes desde direcciones opuestas, y ambas refuerzan la visión de que la ética de la IA no es un problema que deba resolverse una vez y definitivamente, sino una práctica que debe sostenerse.

6

CONCLUSIÓN

Los principales problemas no resueltos de la ética de la IA abarcan dimensiones técnicas, normativas e institucionales. La alineación de valores carece de una formulación canónica de los valores humanos y de un procedimiento legítimo para elegir entre ellos; la investigación sobre equidad está fragmentada entre definiciones incompatibles; la rendición de cuentas está amenazada por modelos opacos y brechas de responsabilidad no resueltas; el diseño de la moral de las máquinas se ve obstaculizado por la heterogeneidad cultural; la gobernanza lucha por operacionalizar y hacer cumplir sus principios; y el campo en su conjunto no ha establecido si son alcanzables soluciones universalmente aceptadas (de Sio & Mecacci, 2021; Jui & Rivas, 2024; Lin, 2021; Prem, 2023; Reuel et al., 2024; Sutrop, 2020). Es poco probable que estos problemas desaparezcan mediante un único avance técnico; más bien, definen un viaje ético continuo e impredecible hacia la garantía de que la IA sirva al florecimiento humano (Stahl et al., 2020). Los análisis críticos examinados aquí refuerzan este veredicto: tanto la crítica radical como la crítica epistemológica interna niegan al campo sus afirmaciones más fuertes, y ambas convergen en la conclusión de que los sistemas de IA deben juzgarse no por una universalidad afirmada sino por una rendición de cuentas demostrable, una validez relativa a propósitos y las decisiones humanas incrustadas en cada representación (Hakim et al., 2026; Lee, 2024; Sowa, 1999; Yampolskiy, 2022).

Traducción completa. Todos los marcadores de cita, cifras (40 millones, 233 jurisdicciones, 172 artículos, 106 y 39 herramien-

tas, 72 cuatrillones, 67,2 %/60,5 %/72,8 %/70,8 %, etc.) y nombres propios se conservan intactos. Si desea que ajuste algún tecnicismo terminológico (por ejemplo, preferir «equidad» frente a «justicia algorítmica», o «rendición de cuentas» frente a «responsabilización»), dígamelo y lo aplico de forma uniforme en todo el texto.

BIBLIOGRAFÍA

Afrikanov, L. (2025). Intuition as a foundational human competence for creativity, ethics, and professional judgment in education in the age of artificial intelligence. *Journal of Education and e-Learning Research*, *12*, 714-719.

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2021). Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management*, *62*, 102433.

Ayling, J., & Chapman, A. (2021). Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics*, *2*, 405-429.

Barenkamp, M., Rebstadt, J., & Thomas, O. (2020). Applications of AI in classical software engineering. *AI Perspectives*, *2*.

Barnett, B. A. (s.f.). 20 Years Later: A Look Back at the Unabomber Manifesto. *9*(6).

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas. *Internet Research*, *33*, 133-167.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *arXiv (Cornell University)*, *352*, 1573-1576.

Brachman, R. J., & Levesque, H. J. (1984). The Tractability of Subsumption in Frame-Based Description Languages. *Proceedings of the National Conference on Artificial Intelligence (AAAI-84)*, 34-37.

Brachman, R. J., & Levesque, H. J. (2004). *Knowledge Representation and Reasoning*. Morgan Kaufmann.

Brändle, C., & Schmidt, M. (2021). Autonomous Driving and Public Reason: a Rawlsian Approach. *Philosophy & Technology*, *34*, 1475-1499.

Cappuccio, M. L., Galliott, J., Eyssel, F., & Lanteri, A. (2023). Autonomous Systems and Technology Resistance: New Tools for Monitoring Acceptance, Trust, and Tolerance. *International Journal of Social Robotics*, *16*, 1-25.

Caton, S., & Haas, C. (2023). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, *56*, 1-38.

Chen, L., Li, Y., Huang, C., Li, B., Xing, Y., Tian, D., Li, L., Hu, Z., Na, X., Li, Z., Teng, S., Lv, C., Wang, J., Cao, D., Zheng, N., & Wang, F.-Y. (2022). Milestones in Autonomous Driving and Intelligent Vehicles: Survey of Surveys. *arXiv (Cornell University)*, *8*, 1046-1056.

Coeckelbergh, M. (2019). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, *26*, 2051-2068.

de Sio, F. S., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, *34*, 1057-1084.

Devedžić, V. (2022). Identity of AI. *Discover Artificial Intelligence*, *2*.

Dignum, V. (2021). The role and challenges of education for responsible AI. *London Review of Education*, *19*.

Dohare, S., Hernandez-Garcia, J., Lan, Q., Rahman, P., Mahmood, A. R., & Sutton, R. S. (2024). Loss of plasticity in deep continual learning. *Nature*, *632*, 768-774.

Dreyfus, H. L. (1972). *What Computers Cant Do*. Harper & Row, Publishers.

Dreyfus, H. L. (1987). From Socrates to Expert Systems: The Limits of Calculative Rationality. *Bulletin of the American Academy of Arts and Sciences*, *40*(4), 15-31.

Dreyfus, H. L., & Dreyfus, S. E. (1986). From Socrates to Expert Systems: The Limits of Calculative Rationality. *Boston Studies in the Philosophy of Science*.

Duch, W. (2007). Intuition, Insight, Imagination and Creativity. *IEEE Computational Intelligence Magazine*, *2*, 40-52.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J. S., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994.

Falk, B. H., & Tsoukalas, G. (2026). The AI Layoff Trap.

Fleming, S. (2021). The Unabomber and the origins of anti-tech radicalism. *Journal of Political Ideologies*, *27*, 207-225.

Fleming, S. (2024). Searching for Ecoterrorism: The Crucial Case of the Unabomber. *American Political Science Review*, *118*, 1986-1999.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*, 411-437.

Galliott, J. (2017, octubre). *The Unabomber on Robots*. Oxford University Press.

Gobet, F., & Chassy, P. (2008). Expertise and Intuition: A Tale of Three Theories. *Minds and Machines*, *19*, 151-180.

Goetze, T. S. (2022). Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 390-400.

Goffi, E. (2021). Escaping the Western Cosm-Ethical Hegemony: The Importance of Cultural Diversity in the Ethical Assessment of Artificial Intelligence. *AI Ethics Journal*, *2*.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*, 99-120.

Hakim, S. A., Fahmi, Z., Siagian, A. H. A. M., Marlina, E., Maryati, I., Apriani, N. F., Saputra, A., Nugroho, C., & Samsudiat, S. (2026, enero). Making AI Outputs Match the Intended Goals (Validity).

Han, S., Kelly, E., Nikou, S., & Svee, E.-O. (2021). Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI & Society*, *37*, 1383-1395.

Henschke, A., & Arora, C. (2024). Pluralism and the Design of Autonomous Vehicles. *Philosophy & Technology*, *37*.

Hernández Servín, J. A. (2026). Intuición como argumentación. En *Horizontes contemporáneos de la filosofía: Ética, Lenguaje y Comunicación*.

Jedličková, A. (2024). Ethical approaches in designing autonomous and intelligent systems: a comprehensive survey towards responsible development. *AI & Society*, *40*, 2703-2716.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *arXiv (Cornell University)*, *1*, 389-399.

Jui, T. D., & Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, *15*, 3095-3125.

Kim, D., & Han, B. (2023, abril). *On the Stability-Plasticity Dilemma of Class-Incremental Learning*.

Kochupillai, M., Lütge, C., & Poszler, F. (2020). Programming Away Human Rights and Responsibilities? "The Moral Machine Experiment" and the Need for a More "Humane" AV Future. *NanoEthics*, *14*, 285-299.

Königs, P. (2022). Artificial intelligence and responsibility gaps: what is the problem? *Ethics and Information Technology*, *24*.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*, 573-616.

Lagarde, R. (2019). A critique of Theodore Kaczynski from an excentric point of view.

Lee, N. (2024, enero). Knowledge Representation and Reasoning.

Leech, G., Garfinkel, S., Yagudin, M., Briand, A., & Zhuravlev, A. (2024, febrero). *Ten Hard Problems in Artificial Intelligence We Must Get Right*.

Lima, G., & Cha, M. (2020, abril). *Responsible AI and Its Stakeholders*.

Lin, C.-F. (2021, octubre). Public Morals, Trade Secrets, and the Dilemma of Regulating Automated Driving Systems. Cambridge University Press.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A. K., & Tang, J. (2022). Trustworthy AI: A Computational Perspective. *ACM Transactions on Intelligent Systems and Technology*, *14*, 1-59.

Martínez-Fernández, S., Bogner, J., Franch, X., Oriol, M., Siebert, J., Trendowicz, A., Vollmer, A. M., & Wagner, S. (2022, abril). *Software Engineering for AI-Based Systems: A Survey* (2).

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the succes-

ses and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419-457.

McKinlay, J., Vos, M. D., Hoffmann, J., & Theodorou, A. (2026). Understanding the Process of Human-AI Value Alignment. *Journal of Artificial Intelligence Research*, *85*.

Mehta, M., Palade, V., & Chatterjee, I. (2023, enero). *Explainable AI: Foundations, Methodologies and Applications*. Springer International Publishing.

Moen, O. M. (2018). The Unabomber's ethics. *Bioethics*, *33*, 223-229.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *26*, 2141-2168.

Munn, L. (2022). The uselessness of AI ethics. *AI and Ethics*, *3*, 869-877.

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. *Archivio istituzionale della ricerca (Alma Mater Studiorum Università di Bologna)*, *39*, 1871-1882.

Ortega-Bolaños, R., Bernal-Salcedo, J., Ortiz, M. G., Sarmiento, J. G., Ruz, G. A., & Tabares-Soto, R. (2024). Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. *Artificial Intelligence Review*, *57*.

Pant, A., Hoda, R., Tantithamthavorn, C., & Turhan, B. (2024). Ethics in AI through the practitioner's view: a grounded theory literature review. *Empirical Software Engineering*, *29*.

Parisi, G. I., Tani, J., Weber, C., & Wermter, S. (2018). Lifelong Learning of Spatiotemporal Representations With Dual-Memory Recurrent Self-Organization. *Frontiers in Neurorobotics*, *12*, 78.

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*, 1-44.

Petersen, P., & Zech, J. (2024, enero). *Mathematical theory of deep learning*. arXiv.

Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. *AI and Ethics*, *3*, 699-716.

Radanliev, P., Santos, O., Brandon-Jones, A., & Joinson, A. (2024). Ethics and responsible AI deployment. *Frontiers in Artificial Intelligence*, *7*, 1377011.

Resnik, D. B., & Hosseini, M. (2024). The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. *AI and Ethics*, *5*, 1499-1521.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P. R., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., … Trager, R. (2024, julio). *Open Problems in Technical AI Governance*.

Rhim, J., Lee, J.-H., Chen, M., & Lim, A. (2021). A Deeper Look at Autonomous Vehicle Ethics: An Integrative Ethical Decision-Making Framework to Explain Moral Pluralism. *Frontiers in Robotics and AI*, *8*, 632394.

Russo, F., Schliesser, E., & Wagemans, J. H. M. (2023). Connecting ethics and epistemology of AI. *AI & Society*, *39*, 1585-1603.

Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information Communication and Ethics in Society*, *19*, 61-86.

Saxe, A., Nelli, S., & Summerfield, C. (2020). If deep learning is the answer, what is the question? *Nature reviews. Neuroscience*, *22*, 55-67.

Schoen, M. P. (2023, noviembre). What Is an Intelligent System? Cambridge University Press.

Servín, J. A. H. (s.f.-a). Intuición como argumentación.

Servín, J. A. H. (s.f.-b). notA SoBre eL hexÁGono LóGico-SeMiótico.

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics. *Journal of Database Management*, *31*, 74-87.

Singh, S. (2025). AI Alignment: Ensuring AI Objectives Match Human Values. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *9*, 1-9.

Souza, L. M., & Castañon, J. A. B. (2021). Public perception of autonomous vehicles: a brief review. *Research Society and Development*, *10*.

Sowa, J. F. (1999). Knowledge Representation: Logical, Philosophical, and Computational Foundations. *CERN Document Server (European Organization for Nuclear Research)*.

Stahl, B. C., Andreou, A. G., Brey, P., Hatzakis, T., Kirichenko, A. S., Macnish, K., Shaelou, S. L., Patel, A., Ryan, M., & Wright, D. (2020). Artificial intelligence for human flourishing – Beyond principles for machine learning. *Journal of Business Research*, *124*, 374-388.

Sutrop, M. (2020). Challenges of Aligning Artificial Intelligence with Human Values. *Acta Baltica Historiae et Philosophiae Scientiarum*, *8*, 54-72.

Tigard, D. W. (2020). There Is No Techno-Responsibility Gap. *Philosophy & Technology*, *34*, 589-607.

Trocin, C., Mikalef, P., Papamitsiou, Z., & Conboy, K. (2021). Responsible AI for Digital Health: a Synthesis and a Research Agenda. *Information Systems Frontiers*, *25*, 2139-2157.

Usher, M., & Barak, M. (2024). Unpacking the role of AI ethics online education for science and engineering students. *International Journal of STEM Education*, *11*.

van de Ven, G. M., Siegelmann, H. T., & Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, *11*, 4069.

Vidal-López, R., González-Ruíz, J. L., Marcial-Romero, J. R., & Hernández-Servín, J. A. (2022). Towards a Formal Language in Law. En D. Valle-Cruz, N. Plata-Cesar & J. L. González-Ruíz (Eds.), *Handbook of Research on Applied Artificial Intelligence and Robotics for Government Processes* (pp. 77-92). IGI Global.

Vidulich, M. A., & Tsang, P. S. (2026). Intuition and deliberation in elite expertise. *Cognitive Research Principles and Implications*, *11*.

Walter, Y. (2024). Managing the race to the moon: Global policy and governance in Artificial Intelligence regulation—A contemporary overview and an analysis of socioeconomic consequences. *Discover Artificial Intelligence*, *4*.

Wang, L., Zhang, X., Li, Q., Zhang, M.-T., Su, H., Zhu, J., & Zhong, Y. (2023). Incorporating neuro-inspired adaptability for continual learning in artificial intelligence. *arXiv (Cornell University)*, *5*, 1356-1368.

Yampolskiy, R. V. (2022). On the Controllability of Artificial Intelligence: An Analysis of Limitations. *Journal of Cyber Security and Mobility*.

Yampolskiy, R. V., & Fox, J. (2012). Safety Engineering for Artificial General Intelligence. *Topoi*.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*.