Mutatis Mutandis: Revista Internacional de Filosofía
Vol. 13 (1) (2026), pp. 1–19, Dossier 1, Art. 2.

Agencia artificial y moralidad delegada: ética y responsabilidad distribuida

Artificial agency and delegated morality: ethics and distributed responsibility)

Martín Bórquez C.
Universidad Adolfo Ibáñez
Región Metropolitana, Chile
martin.borquez@ug.uchile.cl

22 de marzo de 2026

Recibido: 01/08/2025 **Aceptado**: 03/11/2025
DOI: https://doi.org/10.69967/07194773.v13i.609

Resumen

La creciente ubicuidad de la inteligencia artificial (IA) en dominios sociales, económicos y políticos de alta criticidad obliga a examinar los fundamentos éticos de la rendición de cuentas. Los marcos tradicionales, anclados en nociones como la conciencia y la intencionalidad fenoménica, resultan insuficientes para caracterizar sistemas autónomos contemporáneos y mantienen abierta una brecha de responsabilidad en la atribución moral. Bajo este diagnóstico, se sostiene que una ética de la IA conceptualmente plausible y empíricamente informada requiere una definición funcional de la agencia, orientada a criterios observables y desambiguada de la presunción de estados mentales internos. A partir de la ontología informacional de Luciano Floridi, se propone un concepto de agencia artificial entendido como una capacidad interactiva, autónoma y adaptable en la infoesfera. Al articular el análisis filosófico con hallazgos de la psicología moral, se identifican tensiones normativas persistentes y se extraen orientaciones implementables para el diseño y la gobernanza de tecnologías digitales que reontologizan los marcos epistémicos desde los cuales comprendemos la realidad contemporánea.

Palabras clave: Ontología informacional, psicología moral, infoesfera, explicabilidad algorítmica, gobernanza de IA

Abstract

The growing ubiquity of artificial intelligence (AI) in socially, economically, and politically high-stakes domains compels renewed scrutiny of the ethical foundations of accountability. Traditional frameworks, grounded in notions such as consciousness and phenomenal intentionality, prove inadequate for characterizing contemporary autonomous systems and leave open a responsibility gap in moral attribution. On this view, a conceptually plausible and empirically informed AI ethics requires a functional definition of agency—oriented toward observable criteria and disentangled from presumptions about internal mental states. Drawing on Luciano Floridi's informational ontology, this article advances a concept of artificial agency as an interactive, autonomous, and adaptive capacity within the

infosphere. By integrating philosophical analysis with findings from moral psychology, it identifies persistent normative tensions and derives actionable guidance for the design and governance of digital technologies that re-ontologize the epistemic frameworks through which we apprehend contemporary reality.

Keywords: Informational ontology, moral psychology, infosphere, algorithmic explainability, AI governance

1. Introducción: ética para agentes no tradicionales

Hoy resulta difícil describir la vida social sin reconocer la presencia efectiva de sistemas de inteligencia artificial en prácticas ordinarias de comunicación y decisión. Los modelos de lenguaje de gran tamaño inciden en la formulación de textos, en la búsqueda de información y en la elaboración de respuestas que muchas veces se toman como insumos epistémicos, con impacto directo en la producción y circulación del conocimiento (Bender et al., 2021; Danaher, 2024). De modo análogo, algoritmos especializados automatizan o condicionan decisiones en finanzas, medicina y logística, mediante procedimientos de predicción, clasificación y asignación de recursos (O'Neil, 2016; Taddeo & Floridi, 2018). En este contexto, la IA deja de ser una promesa futurista y adquiere el estatuto de una fuerza estructurante de nuestro tiempo, precisamente porque altera condiciones de acción y criterios de justificación en dominios de alta criticidad. Su ubicuidad plantea desafíos éticos de elevada complejidad que, con frecuencia, desbordan los marcos normativos vigentes, dado que estos fueron concebidos para escenarios donde la agencia, la intención y la trazabilidad se modelan a escala humana, acumulando tensiones morales constitutivas de un giro paradigmático (Coeckelbergh, 2020a; Crawford, 2021). Ante un daño provocado por un sistema autónomo, la pregunta por la atribución moral se vuelve ineludible. ¿Debe responder el usuario, el propietario, la empresa que lo diseñó y lo desplegó, o cabe imputar responsabilidad, en algún sentido, al propio sistema? Esta dificultad remite a lo que Andreas Matthias (2004) denominó brecha de responsabilidad (*responsibility gap*), entendida como un vacío normativo que surge cuando los efectos de agentes artificiales autónomos no pueden atribuirse con justicia y trazabilidad a responsables identificables, dentro de cadenas verificables de decisión. La raíz de esta brecha reside en marcos éticos diseñados para agentes humanos ideales, dotados de conciencia, deseos e intenciones. Los sistemas actuales carecen de esos rasgos en sentido biológico y fenoménico, lo que vuelve frágiles las inferencias habituales sobre imputación moral (Epstein, 2016). Pues bien, una vez reconocido ese límite, el problema se desplaza hacia el plano técnico-normativo de la imputación, esto es, hacia los dispositivos institucionales mediante los cuales se ordenan riesgos, deberes y compensaciones. En el derecho positivo, la personalidad jurídica constituye una herramienta técnica orientada a ordenar patrimonios, riesgos y deberes. Su eventual extensión a sistemas artificiales podría facilitar reglas de imputación y esquemas de garantía, sin presuponer por ello una agencia moral fuerte, aunque exigiría criterios adicionales para abordar la responsabilidad en sentido sustantivo (Ruiz Osuna, 2025). En esta misma línea, elevar la exigencia de un "fantasma en la máquina" a condición de imputación supone, *prima facie*, un error categorial que esclerotiza el debate y amplifica la exposición a riesgos técnicos concretos (Sinnott-Armstrong & Skorburg, 2021). A la luz de ello, conviene desplazar la atención hacia las mediaciones técnicas que organizan la per-

cepción, la acción y la evaluación moral. En la filosofía de la tecnología, autores como Verbeek (2006, 2011) han mostrado que los artefactos técnicos funcionan como mediaciones activas. Intervienen en la forma en que percibimos, actuamos y atribuimos valor moral al mundo. Sobre esta base, la *machine ethics* aspira a una ética funcional, entendida como conjuntos implementables de reglas, políticas y normas. El propósito consiste en orientar decisiones y evaluar comportamientos mediante criterios auditables, transparentes y explicables. En este trabajo, tales formas se entienden como expresiones de un desempeño normativo bajo control humano, en sintonía con modelos de moralidad distribuida y con la noción de *faultless responsibility* (Allen et al., 2000; Floridi, 2016; Wallach & Allen, 2008). La inteligencia artificial contemporánea profundiza esta línea de análisis. Pasa de desempeñar un rol predominantemente instrumental a constituirse como una presencia activa en el ecosistema social. Con ello, introduce modalidades de agencia *sui generis* que desplazan los contornos de la acción y vuelven más problemática la imputación moral.

El trabajo se organiza en torno a una tesis central. Para desarrollar una ética de la inteligencia artificial con coherencia conceptual y aplicabilidad normativa, resulta necesario dejar atrás la exigencia de conciencia e intencionalidad como condiciones *sine qua non* de una agencia moralmente significativa. En su lugar, se propone una concepción funcional e informacional de la agencia, formulada en el marco de la filosofía de la información (Floridi & Sanders, 2004).

En función de lo anterior, la agencia se define como la capacidad de una entidad para relacionarse con su entorno, producir cambios en él y ajustar su conducta en función de información o estímulos, con grados variables de autonomía, adaptabilidad y teleonomía, entendida como orientación operativa hacia fines (Floridi, 2025). Este giro conceptual, alineado con el modo en que la IA contemporánea efectivamente funciona, habilita un marco normativo para pensar la moralidad delegada (*delegated morality*) y la responsabilidad sin reproche (*faultless responsibility*) (Floridi, 2016). Ambas pueden comprenderse como modalidades distribuidas de agencia moral que se despliegan en la infoesfera.

Para sostener esta tesis, el trabajo adopta un enfoque deductivo. La segunda sección examina críticamente la polisemia del concepto de "inteligencia" en IA y distingue entre el horizonte de una Inteligencia Artificial General (AGI) y el panorama actual, marcado por una inteligencia de carácter reproductivo e ingenieril. La tercera sección establece las bases de una ontología informacional. Allí se define la agencia artificial en términos funcionales y se la sitúa dentro del ecosistema de la infoesfera y de las tecnologías de tercer orden. Luego se aborda la moralidad delegada y la brecha de responsabilidad, con respuestas apoyadas en una concepción distribuida de la agencia moral. A continuación, la discusión filosófica se articula con hallazgos recientes de la psicología moral empírica. Se muestra cómo las intuiciones y expectativas humanas frente a la IA inciden en el diseño y la gobernanza de sus marcos éticos. El cierre propone un enfoque híbrido que integra principios, prácticas y derechos orientados a resolver dilemas éticos concretos, junto con una comprensión contextual y empíricamente informada de la interacción humano–máquina. Con ello, se presenta una propuesta de gobernanza teóricamente coherente y susceptible de implementación frente a los desafíos que plantea la IA en la sociedad contemporánea.

2. Qué cuenta como inteligencia en IA: disputas conceptuales y límites epistémicos

El sintagma *inteligencia artificial* arrastra una ambigüedad semántica y conceptual ampliamente reconocida en la literatura especializada. El adjetivo "artificial" abre cuestiones ontológicas sugestivas, aunque el núcleo de la indeterminación se concentra en el sustantivo "inteligencia", cuya delimitación teórica y criterios de aplicación permanecen disputados. La dificultad para ofrecer una caracterización unívoca de la IA expresa, en el fondo, un desacuerdo epistemológico sobre qué cuenta como inteligencia, qué dimensiones son constitutivas y qué indicadores permiten atribuirla con validez. Shane Legg y Marcus Hutter (2007) abordaron este punto de manera empírica en un estudio seminal, al sistematizar más de setenta definiciones provenientes de la psicología, la filosofía y las ciencias de la computación, mostrando la fragmentación disciplinar que subyace al concepto. Ese mapa exhibe un continuo que va desde concepciones centradas en aprendizaje, adaptación y desempeño en entornos, hasta enfoques que privilegian razonamiento abstracto, resolución de problemas y comprensión simbólica del lenguaje. Esta dispersión ha obstaculizado la consolidación conceptual del campo, en la medida en que incentiva atribuciones antropomórficas a sistemas algorítmicos. El resultado es una confusión categorial entre desempeño computacional y capacidades mentales, junto con un desplazamiento de los límites epistémicos que impone la arquitectura informacional de tales sistemas, límites que quedan subteorizados precisamente cuando el vocabulario psicológico sustituye a la descripción funcional (Nyholm, 2020).

En otra dirección, el proyecto fundacional de Dartmouth formuló la tesis de que "cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, describirse con tal precisión que una máquina podría simularla" (citado en Bringsjord & Govindarajulu, 2020). Esta idea, reforzada por hitos como el Test de Turing, consolidó una comprensión de la IA que vincula la cognición con su formalización computacional y su eventual reproducción maquínica. De esa tradición surge la distinción canónica entre IA Fuerte o Inteligencia Artificial General (AGI), orientada a construir una máquina con capacidad intelectual general comparable o superior a la humana (Bostrom, 2014), e IA Débil o Estrecha (*Narrow AI*), centrada en desempeños acotados a tareas específicas. Dentro del ideal de la AGI suelen distinguirse dos vertientes. Una define la inteligencia general como competencia funcional capaz de resolver problemas en múltiples dominios. La otra la concibe como emulación de los procesos cognitivos humanos, más cercana al programa explicativo de las ciencias cognitivas contemporáneas (Floridi, 2024).

Con todo, la narrativa de la simulación cognitiva descansa en una metáfora epistemológicamente problemática, la del cerebro como "procesador de información" o "computador biológico". Robert Epstein (2016), en *The Empty Brain*, advierte que la anfibología aparece cuando una metáfora funcional se toma como descripción literal. Un computador digital manipula símbolos discretos según reglas explícitas. El cerebro, en cambio, se sostiene en dinámicas biofísicas que organizan actividad neuronal y flujos de energía en patrones relativamente estables, sensibles al entorno y a la historia del organismo. Esos patrones hacen posible la memoria, la cognición y la conducta sin requerir la idea de algoritmos predefinidos en sentido puramente

técnico. En este registro, hablar de "procesamiento" designa un fenómeno biológico y dinámico, más que un procedimiento formal y simbólico propio de los sistemas computacionales.

En última instancia, la metáfora del procesamiento de información puede conservar valor heurístico, pero no alcanza el estatuto de una descripción literal del funcionamiento neurobiológico. A la manera de los enfoques enactivo y autopoiético, lo decisivo consiste en situar la actividad cerebral en una economía de acoplamiento cuerpo–entorno, como una realidad encarnada, plástica y autoorganizada (Cisek, 1999; Varela, 1990). La cognición se comprende, así, como una dinámica de interacción y regulación continua entre organismo y medio, antes que como una manipulación interna de representaciones simbólicas en el sentido del cognitivismo clásico (Varela, 1990).

Esta crítica al representacionalismo permite volver sobre el marco del cognitivismo histórico. Corbí y Prades (1999) observan que dicho paradigma concibió la mente como un sistema de procesamiento simbólico, en el cual los estados mentales se asimilaban a representaciones físicas, subpersonales, manipuladas mediante reglas sintácticas. El rendimiento de este enfoque fue real, en particular al formalizar ciertos vínculos entre mente y computación. A la vez, consolidó una metáfora influyente, la del cerebro como computador, que carece de correlato neurobiológico directo. Los desarrollos ulteriores, entre ellos el conexionismo y diversas teorías dinámicas postcognitivistas, tensionaron ese supuesto. Mostraron que la actividad mental puede comprenderse mejor como emergencia de redes distribuidas y procesos autoorganizativos que no se dejan reducir a reglas simbólicas explícitas (Rabossi, 1995; Varela, 1990). En contraste, las máquinas sí procesan información en sentido literal. Trabajan con símbolos discretos y bits cuya manipulación depende de reglas sintácticas determinadas (Corbí y Prades, 1999). La confusión entre estos planos, el biológico y el computacional, favorece atribuciones indebidas a la IA, tales como pensamiento o comprensión. Mitchell y Krakauer (2023) advierten que, aunque estos sistemas alcancen rendimientos a veces notables, sus transformaciones internas no constituyen garantía alguna de intencionalidad ni de contenido semántico propio. Precisar la diferencia resulta, por ello, decisivo. Evita el error categorial que equipara cognición biológica y computación simbólica, permitiendo, sobre esa base, estabilizar un marco conceptual aséptico para el análisis de la agencia artificial.

Ante la polisemia que envuelve el sintagma "inteligencia artificial", Luciano Floridi (2024) propone una distinción taxonómica más precisa y teóricamente productiva. Con ello, el problema se desplaza desde la simulación de la cognición hacia la caracterización de la capacidad funcional del sistema, concebida como un repertorio de competencias operacionalizables. Esta orientación converge con definiciones recientes de inteligencia general artificial (Hendrycks et al., 2025), que enfatizan la distancia entre la versatilidad cognitiva humana, anclada en trayectorias evolutivas, y las modalidades de desempeño computacional propias de arquitecturas algorítmicas. Sobre esa base, Floridi distingue dos grandes vertientes. La primera, la IA productiva y cognitiva, prolonga el proyecto clásico de la Artificial General Intelligence (AGI). Su horizonte apunta a una inteligencia análoga a la humana, ya sea mediante la construcción de artefactos con capacidades generales o bien, mediante el modelamiento de la cognición humana con fines explicativos (Bostrom, 2014). En el estado actual,

ese programa se mantiene predominantemente en un registro teórico y, en buena medida, especulativo (Floridi, 2024). La segunda vertiente, en cambio, la IA reproductiva y de ingeniería, se identifica con la IA estrecha que estructura el panorama contemporáneo. Su orientación consiste en reproducir desempeños asociados a conductas inteligentes en dominios acotados, como una disciplina de diseño que construye sistemas capaces de ejecutar funciones que, en humanos, suelen requerir inteligencia. A partir de esta distinción, los sistemas que hoy concentran atención y preocupación, entre ellos los modelos de lenguaje de gran tamaño y los sistemas de visión computacional, se ubican inequívocamente en la segunda categoría. En la formulación de Bender y sus colegas (2021), se trata de "loros estocásticos" (*stochastic parrots*). Es decir, arquitecturas altamente eficientes para anticipar secuencias lingüísticas probables a partir de patrones estadísticos distributivos extraídos de volúmenes masivos de datos, pero con un desacople persistente respecto de un anclaje semántico y de una comprensión objetiva del mundo al que, en apariencia, remiten sus enunciados formales. En este registro, cobra fuerza la reinterpretación floridiana de la fórmula de Clausewitz. A saber, la IA reproductiva aparece como continuación del comportamiento inteligente por otros medios, a través de simulación algorítmica carente de intención y de conciencia fenoménica del mundo (Floridi, 2024). El énfasis recae en la externalización de productos del trabajo cognitivo humano, lo que habilita una forma de amplificación cognitiva cualitativamente distinta de las disponibles hasta ahora. Esa potencia reproductiva se encuentra condicionada por la composición de los datos de entrenamiento. Cuando estos son históricos, se derivan del registro digital de la actividad humana, arrastrando sesgos estructurales, con efectos capaces de cristalizar desigualdades sociales, en línea con las llamadas "armas de destrucción matemática" (O'Neil, 2016). Además, los datos también pueden ser sintéticos. Esto es, generados en entornos controlados para modelar situaciones específicas, o híbridos, cuando integran información histórica y sintética en un mismo proceso de entrenamiento. A ello se suma una condición material frecuentemente invisibilizada. La construcción de los conjuntos de datos suele descansar en contingentes amplios de trabajo humano precarizado, los "turcos mecánicos" del siglo XXI, responsables de etiquetar, depurar y validar información, lo que expone el sustrato humano que sostiene la apariencia de autonomía de estos sistemas (Crawford, 2021). Con todo, comprender la IA contemporánea como una ciencia de lo artificial en sentido ingenieril, orientada a la reproducción funcional de conductas inteligentes, permite asentar una ética apoyada en evidencia y en descripciones técnicas de capacidades, en lugar de proyecciones antropomórficas. El debate, entonces, se reubica. Deja de girar en torno a conjeturas sobre conciencia artificial y pasa a concentrarse en cadenas de producción, dinámicas laborales e implicaciones sociales. Esta reorientación constituye una condición necesaria para someter la tecnología a un escrutinio informado y así encauzar su desarrollo conforme a exigencias de explicabilidad, transparencia y responsabilidad.

3. Hacia una ontología informacional de la agencia

Una vez analizada la ambigüedad semántica del concepto de inteligencia, el siguiente paso consiste en establecer el concepto de agencia sobre un terreno epistémico claro y acorde a la naturaleza específica de los sistemas de inteligencia artificial. Si tales

sistemas carecen de pensamiento en sentido fenoménico y aun así ejecutan acciones con efectos morales concretos, se vuelve necesario un marco ontológico capaz de describir ese hacer sin proyectar, por defecto, estados mentales.

Este marco se deriva de la filosofía de la información, en particular de la ontología informacional desarrollada por Luciano Floridi (2014). Su premisa central propone comprender la realidad contemporánea como una *infoesfera*, un entorno global de información, tanto digital como analógica, constituido por el conjunto de entidades, procesos y propiedades informacionales. En ese medio coexisten, bajo un mismo horizonte descriptivo, agentes naturales junto a agentes artificiales (Almendros & Echeverría, 2019; Floridi, 2025).

Notablemente, en *The Fourth Revolution* (2014), Floridi sostiene que las tecnologías digitales han impulsado una revolución ontológica *stricto sensu*. La *infoesfera* se comprende como una capa informacional integrada a la realidad material, capaz de transformar los modos ordinarios de experiencia. De ahí su noción de *onlife*, una forma de existencia híbrida en la que la distinción entre lo *online* y lo *offline* adquiere un carácter estructuralmente pervasivo.

En el presente, marcado por una reontologización asociada a la integración extendida de sistemas digitales, las tecnologías de la información pueden describirse como tecnologías de primer orden. En esa condición, intervienen en la determinación de estados del mundo, desplazando gradualmente a los seres humanos hacia funciones de supervisión o hacia el lugar de beneficiarios externos al circuito de control directo, en lo que figurativamente se denomina *out of the loop*.

Bajo esta lectura, la inteligencia artificial se sitúa dentro de lo que Floridi (2013) denomina tecnologías de tercer orden. Las tecnologías de primer orden median la relación entre el ser humano y la naturaleza, como ocurre con una caña de pescar que amplía la capacidad de acción. En cambio, las tecnologías de segundo orden median la relación entre una tecnología y la naturaleza, como un termostato que regula una caldera. Las tecnologías de tercer orden se distinguen de sus predecesoras por su capacidad de interactuar entre sí, prescindiendo de intervención humana directa. Un ejemplo ilustrativo es una plataforma logística automatizada, en la que un algoritmo gestiona el despacho de productos, coordina vehículos autónomos para el transporte y comunica a sensores de bodega la reposición de inventario, todo ello sin intervención humana en tiempo real. Con tecnologías de este tipo, la distinción entre herramienta y entorno pierde nitidez. El sistema técnico deja de presentarse como instrumento y pasa a constituir un entorno activo, un medio con el que interactuamos y que, a la vez, nos permea en múltiples niveles (Floridi, 2007).

En el marco de estas dinámicas interaccionales se vuelve necesario precisar el concepto de agencia. Floridi y Sanders (2004) proponen una definición de agente informacional apoyada en criterios funcionales y observables, al margen de nociones mentales como conciencia, creencias o deseos. Su propuesta identifica tres propiedades: Interactividad, autonomía y adaptabilidad. En primer lugar, la interactividad remite a la influencia recíproca entre el agente y su entorno. Un sistema de IA recibe datos de entrada y produce decisiones o acciones que alteran el estado del mundo en el que se inserta. La autonomía alude a la capacidad de modificar su estado interno, desplegando cursos de acción sin intervención externa inmediata, como ocurre con

algoritmos de *trading* automatizado que ejecutan transacciones sin supervisión humana en tiempo real. La adaptabilidad designa el ajuste de los patrones de interacción y de las reglas de transición de estado a partir de la experiencia, con el aprendizaje por refuerzo como caso paradigmático. Un sistema que satisface estas tres condiciones puede considerarse un agente artificial en sentido pleno, con alcance conceptual más allá del uso meramente metafórico (Floridi, 2024). Por su generalidad, este marco abarca, *mutatis mutandis*, agentes biológicos, como seres humanos o animales, agentes artificiales, como vehículos autónomos o sistemas de recomendación, además de formas de agencia colectiva, como una red eléctrica inteligente. En este último caso, la capacidad de acción emerge de una organización distribuida. La atribución de responsabilidad moral adquiere, por tanto, rasgos propios que exceden la suma de componentes. Detengámonos brevemente en este punto. Concebir la agencia en clave funcional ofrece, entonces, el soporte teórico central para una ética de la IA empíricamente informada y normativamente mensurable. Bajo esta perspectiva, un sistema de IA puede ser moralmente significativo como *Artificial Moral Agent* (AMA), en la formulación temprana de Allen, Varner y Zinser (2000), en virtud de su condición de fuente autónoma de acciones con consecuencias moralmente significativas. La importancia ética de tales acciones estriba en sus efectos sobre el mundo, en especial cuando inciden en el bienestar o el perjuicio de agentes morales humanos. En efecto, al desvincular la noción de agencia de una arquitectura cognitiva interna—esto es, de la necesidad de creencias, conciencia o deseos—, se abre la posibilidad de repensar la responsabilidad moral desde parámetros más acordes a la naturaleza distribuida, opaca y no intencional de muchos sistemas tecnológicos actuales.

El giro propuesto tiene una consecuencia práctica decisiva. En vez de exigir que la IA se acomode a un entorno humano ambiguo y contingente —tarea que exigiría algo cercano a la AGI—, la tendencia dominante invierte la dirección. El entorno se ajusta, *de facto*, para calzar con las capacidades topológicas de la IA. Esta inversión se vuelve visible en la tensión entre reglas restrictivas, que delimitan qué puede hacer el sistema y reglas constitutivas, que definen el *locus* en el que la IA puede operar. Un almacén de Amazon, con códigos QR en el suelo y estanterías estandarizadas, ilustra bien el principio. El espacio se diseña para que los robots Kiva lo recorran con fiabilidad. En términos concretos, el desplazamiento del robot se vuelve simple porque el espacio logístico queda ordenado en función de sus sensores y rutinas, de modo que la tarea puede ejecutarse mediante instrucciones sensoriomotoras básicas. Entonces, a medida que se delegan funciones cada vez más complejas a agentes de inteligencia estrecha, esta estrategia se encadena y se amplifica. Sus efectos tienen alcance ontológico, porque da forma al mundo físico, social y digital para volverlo legible y manejable por estos nuevos agentes (Danaher & Sætra, 2022). Al moldear el entorno para que resulte accesible a sistemas artificiales, también cambia el horizonte de lo que puede ser percibido, conocido y habitado. Mientras los organismos biológicos ajustan su conducta mediante procesos adaptativos internos, los sistemas artificiales dependen de una adaptación externa, lograda por la disposición del entorno. Esta inversión del principio adaptativo constituye una de las transformaciones epistémicas más sugestivas y, a la vez, menos advertidas que la IA está introduciendo en nuestra relación moral con el mundo.

4. La brecha de responsabilidad y el desafío de la moralidad delegada

La inteligencia artificial puede entenderse como una forma de agencia funcional, interactiva y adaptativa, con una realidad ontológica definida por sus efectos. Esta caracterización ilumina uno de los dilemas éticos más apremiantes del campo, la llamada *brecha de responsabilidad* (Matthias, 2004). Cuando un sistema carente de estados mentales, intencionalidad o culpa ocasiona un perjuicio, los marcos clásicos del derecho civil exhiben sus límites, pues se sustentan en nociones de culpa, previsibilidad e intención, además de deberes de cuidado y de la posición de garante. En estas condiciones, una salida jurídica prudente consiste en tratar a tales sistemas como un *tertium genus*, un estatus intermedio entre personas y cosas con efectos patrimoniales. Ello exige ajustar los criterios de imputación para contemplar esquemas de responsabilidad objetiva, sin culpa, frente a riesgos normativos asociados al despliegue de sistemas autónomos (Ruiz Osuna, 2025). Más que una disputa escolástica de positivismo jurídico, el debate apunta a una actualización normativa de la atribución entre personas y agentes artificiales. De ahí la urgencia de instituir mecanismos de trazabilidad, auditoría y compensación, con coordinación entre Estado, mercado y sociedad civil. El objetivo es doble. Por una parte, conciliar equidad con eficiencia. Por otra, habilitar formas renovadas de rendición de cuentas en arreglos de decisión donde la acción humana se entrelaza con salidas algorítmicas, bajo intervención de sistemas autónomos.

La preocupación expuesta se funda en los efectos estructurales que la integración de estos sistemas introduce en la organización de nuestras sociedades. John Danaher (2016) denomina a esta tendencia *algocracia*, un modelo de gobierno en el que un volumen creciente de decisiones con implicaciones morales se delega en sistemas automatizados. Este desplazamiento de la agencia se vincula con lo que Wallach y Allen (2008) conceptualizan como moralidad delegada, entendida como la cesión deliberada de juicios y acciones éticas a agentes artificiales. En la práctica, muchas veces bajo esquemas de decisión asistida y con grados variables de supervisión humana, se delegan decisiones que inciden directamente en la vida de las personas, tales como la evaluación crediticia o la asignación clínica de recursos, además de dominios especialmente sensibles como el *targeting* militar. En este último ámbito resulta imperativo sostener un control humano significativo (Sparrow, 2007), con el fin de resguardar la responsabilidad moral y la trazabilidad de decisiones potencialmente letales. Cada acto de delegación presupone confianza en que el sistema se alineará con ciertos valores, aunque simultáneamente difumina *ipso facto* las fronteras tradicionales de la responsabilidad y la rendición de cuentas.

Superar esta brecha exige evitar dos atajos igualmente estériles. Uno consiste en antropomorfizar al agente artificial para adjudicarle culpa en sentido humano. El otro consiste en diluir la responsabilidad humana mediante una apelación acrítica a la "autonomía" del sistema. El primer gesto incurre en un error categorial. El segundo equivale a una abdicación moral. La salida más fértil, como propone Floridi (2016), pasa por actualizar el propio concepto de responsabilidad. En ese marco adquiere centralidad la noción de responsabilidad irreprochable (*faultless responsibility*), entendida como una atribución moral anclada en la relación causal entre un agente

y un resultado éticamente concreto, con independencia de intención o culpabilidad. Esta objetivación de la responsabilidad evita reificar al agente en destinatario de un juicio moral, aunque exige asumir consecuencias y activar mecanismos de respuesta frente a los efectos producidos. *In lato sensu*, atribuir responsabilidad irreprochable a un agente artificial equivale a identificarlo como nodo causal clave en la cadena de eventos. Tal reconocimiento no habilita modulaciones punitivas como castigo o culpa, prácticas carentes de *sensus communis* en este dominio. Más bien, funciona como un paso preliminar e indispensable para un segundo nivel de análisis, el de la responsabilidad distribuida (Coeckelbergh, 2020b). Conforme a esta lectura, si el agente artificial figura como responsable irreprochable del evento, la responsabilidad imputable, aquella que activa deberes de reparación, compensación o modificación, se distribuye a lo largo de la cadena causal que hizo posible su diseño, despliegue y desempeño. Esa trama abarca equipos de diseño e ingeniería, cuyos márgenes de decisión suelen quedar definidos por mandatos organizacionales, a las corporaciones que lo comercializaron, a los organismos encargados de regular su uso y, finalmente, a los usuarios que lo incorporaron en sus prácticas. En esta clave analítica, la brecha de responsabilidad deja de entenderse como un vacío que reclama un único culpable. Se presenta, más bien, como un espacio complejo y cartografiable, que habilita la atribución gradual y tipológicamente diferenciada de responsabilidad a múltiples actores dentro de una misma cadena moral de causalidades (Mittelstadt et al., 2016). Un ejemplo permite fijar la distinción. Si un vehículo autónomo ocasiona un accidente, el sistema puede identificarse como causa proximal del evento, en calidad de responsable irreprochable. La responsabilidad imputable se distribuye, en cambio, por niveles. Puede recaer en la empresa responsable del producto por fallos en los sensores (responsabilidad por diseño), al propietario por omitir las actualizaciones de software (responsabilidad por mantenimiento), e incluso a las autoridades públicas por autorizar la circulación de una tecnología aún inmadura (responsabilidad por regulación).

Obsérvese conforme a lo indicado que el modelo de responsabilidad distribuida exige una condición técnico-epistémica decisiva, la explicabilidad (*explainability*). Para trazar responsabilidades a lo largo de la cadena de causas y efectos que desemboca en una decisión automatizada, se requiere comprender por qué un sistema actuó de determinada manera y, sobre esa base, determinar a quién corresponde la responsabilidad ética por sus consecuencias. En este punto conviene ser precisos. Los sistemas descritos como “caja negra”, debido a la opacidad de sus dinámicas internas, obstaculizan la atribución de responsabilidad incluso para quienes los desarrollan. Esa opacidad rompe el nexo inteligible entre agencia técnica y *accountability* humana (Matthias, 2004). Frente a este desafío, la Inteligencia Artificial Explicable (*Explainable AI*, XAI) se presenta como un objetivo técnico prioritario a la vez que como un imperativo ético con efectos sociales contrastables (Coeckelbergh, 2020b). Su propósito consiste en aumentar la inteligibilidad de sistemas opacos, de modo que resulte posible reconstruir la cadena causal entre decisiones y resultados. En este sentido, la explicabilidad se vuelve condición previa para la rendición de cuentas. Con todo, incluso cuando se recupera un grado relevante de inteligibilidad, se impone una cautela adicional. Toda decisión basada en datos y modelos formales enfrenta, *ceteris paribus*, límites constitutivos, dado que la complejidad del mundo excede la capacidad representacional de cualquier esquema algorítmico, por transparente que resulte (Taddeo

& Floridi, 2005). Por ello, la aspiración a una IA ética, incluso cuando se apoya en XAI, pierde eficacia si carece de un andamiaje institucional que asegure mecanismos exigibles de rendición de cuentas, junto con estándares técnico-normativos capaces de sostener la aplicabilidad contextual de las decisiones automatizadas (Mittelstadt, 2019).

5. Psicología moral de la interacción humano–IA: expectativas, confianza y atribución normativa

La robustez de un marco normativo en ética de la inteligencia artificial depende tanto de la coherencia de sus fundamentos filosóficos como de su capacidad para articularse con la dimensión empírica de la interacción entre agentes humanos y artificiales. Para evitar que los principios normativos se transformen en construcciones meramente formales, desconectadas de la práctica, resulta necesario anclarlos en una comprensión sistemática y empíricamente informada de los modos en que las personas perciben, valoran y otorgan confianza a agentes artificiales.

En este punto, la psicología moral, con su orientación empírico-conductual, ofrece una contribución relevante al debate normativo. Permite identificar tensiones, anticipar respuestas sociales y orientar un diseño ético que aspire a plausibilidad psicológica y sostenibilidad social. Aun así, las preferencias observadas empíricamente deben distinguirse con rigor de la justificación normativa de los principios éticos, cuyo fundamento es argumentativo antes que estadístico. Pues bien, un ejemplo paradigmático de esta línea es el experimento *Moral Machine*, desarrollado por Awad et al. (2018), que mediante una plataforma global recopiló cerca de cuarenta millones de juicios sobre dilemas morales de tipo tranvía aplicados al contexto de los vehículos autónomos.

Los resultados evidenciaron constantes morales relativamente estables, entre ellas la atribución de un estatus diferenciado a la vida humana y la preferencia por minimizar tanto el daño como el número de víctimas. La muestra estuvo condicionada por brechas de acceso digital y por sesgos en la distribución territorial. Esa restricción exige interpretar el hallazgo en términos socioculturales y tratarlo como un patrón situado. Las regularidades identificadas coexisten con variaciones sustantivas y dependen, en gran medida, del contexto indexical en el que se formulan los juicios morales. En esa línea, las decisiones de los participantes se asociaron a variables como la edad, el género y el estatus social percibido de las potenciales víctimas. El conjunto sugiere que los juicios morales se elaboran dentro de coordenadas sociocognitivas específicas, quedando expuestos a sesgos heterogéneos. En ello reside la importancia del estudio, que aporta bases empíricas para sostener que la moralidad delegada no se deja reducir a una regla única y adopta, más bien, una valencia de carácter indexical. Empero, es bien sabido que los dilemas tipo tranvía han recibido críticas por su simplificación, ya que tienden a reducir la ética a un cálculo de víctimas, dejando fuera la complejidad ordinaria de la gestión del riesgo. Su valor experimental, aun así, radica en aislar tensiones estructurales del juicio moral y volverlas fácticamente analizables.

A partir de estas consideraciones**,** la heterogeneidad de las preferencias exige, por tanto, un enfoque analítico atento al alcance inferencial de los datos y capaz de se-

guir cómo, en contextos concretos, las intuiciones morales median la tensión entre utilitarismo y deontología, especialmente cuando esos principios se aplican a agentes no humanos.

Una de las líneas más fecundas de este enfoque empírico examina la tensión entre los dos grandes paradigmas de la ética normativa, el utilitarismo, orientado a maximizar bienestar agregado, y la deontología, centrada en deberes y restricciones morales. Un hallazgo recurrente, además contraintuitivo, muestra expectativas asimétricas hacia agentes humanos frente a agentes artificiales. Myers y Everett (2025) indican que muchas personas esperan de asesores morales artificiales una orientación más utilitarista que la atribuida a consejeros humanos, pues la IA suele asociarse con cálculo racional, imparcialidad y eficiencia. En escenarios experimentales, esa asociación la vuelve una opción aparentemente idónea cuando el dilema exige ponderar bienes comunes o evaluar consecuencias distributivas. Con todo, el mismo resultado abre una paradoja axiológica para el diseño ético. Myers y Everett (2025) muestran que, aunque se espera un sesgo utilitarista en la IA, la confianza disminuye cuando asesores humanos o artificiales explicitan razonamientos de ese tipo. Un agente artificial que declara disposición a sacrificar a una persona por un bien mayor tiende a percibirse como menos confiable y moralmente problemático. De ahí el dilema aporético de diseño. Cabe alinear, entonces, el sistema con expectativas públicas de imparcialidad utilitarista, asumiendo el riesgo de una percepción de frialdad moral, o bien incorporar restricciones deontológicas, como la prohibición de causar daño intencionalmente, para favorecer aceptación social, aun cuando ello pueda conducir en ciertos casos a desenlaces moralmente subóptimos. En consecuencia, esta tensión pone de relieve que la optimización algorítmica no garantiza, por sí misma, la aceptabilidad social. Gigerenzer (2022) subraya que, en entornos complejos e inciertos, la inteligencia humana basada en heurísticas rápidas y frugales puede superar a algoritmos sofisticados que dependen de grandes volúmenes de datos y exhiben fragilidad ante situaciones imprevistas. La salida ética, por tanto, apunta al diseño de arreglos de colaboración humano–máquina que integren fortalezas complementarias, en lugar de perseguir la construcción de una "máquina perfectamente moral" (Sinnott-Armstrong & Skorburg, 2021; Taddeo & Floridi, 2018).

Finalmente, la psicología de la interacción con la inteligencia artificial identifica riesgos cognitivos que exigen una gestión cualitativa (Cave et al., 2019). Entre los más significativos figura la complacencia moral, cercana a formas de sesgo de confirmación, mediante la cual las personas tienden a aceptar recomendaciones de sistemas automatizados sin un examen crítico suficiente, con el consiguiente desplazamiento del propio juicio ético. Kaas (2024) sitúa este fenómeno dentro de una "tormenta tecnológica perfecta", en la que la opacidad del sistema, la autoridad epistémica atribuida y la promesa de eficiencia convergen, erosionando gradualmente la capacidad humana de discernimiento moral. Desde este ángulo, una ética aplicada de la IA ha de considerar tanto la moralidad inscrita en el artefacto como el modo en que su arquitectura cognitiva incide sobre la moralidad del juicio humano que interactúa con él. Aquí resulta pertinente la noción de moralidad materializada desarrollada por Verbeek (2006, 2011), según la cual la tecnología media la acción, codifica formas de comprensión y orienta decisiones que repercuten en prácticas morales de la vida pública y de la vida privada. Por consiguiente, la evidencia empírica amplía el alcance del razonamiento normativo al situarlo en contextos concretos, lo que permite dilucidar

sesgos cognitivos y normativos que la gobernanza ética debe reconocer, interpretar y atender.

6. Gobernanza normativa de la agencia artificial: principios, prácticas y derechos

La transición desde el análisis filosófico y psicológico hacia una propuesta de gobernanza concreta constituye el tramo final, quizá el más decisivo, para una ética de la inteligencia artificial que aspire a incidir efectivamente en la sociedad contemporánea. Un marco normativo que reconozca la agencia funcional, la responsabilidad distribuida y las complejidades de la psicología moral humana ha de traducirse en principios, prácticas y, eventualmente, nuevos derechos. Su propósito es orientar el desarrollo e implementación de tecnologías paradigmáticas que modulan las condiciones del juicio moral y de la acción humana. Se trata, por tanto, de un ejercicio pragmático, dirigido a identificar riesgos, robustecer la confianza pública, direccionando la inteligencia artificial hacia fines normativamente deseables. Entre esos fines destaca su proyección como fuerza orientada al bien común (Taddeo & Floridi, 2018). Esa aspiración exige mantener a la vista usos perjudiciales que ya adquieren densidad social y política dentro de la infoesfera, como la manipulación informativa mediante *deepfakes*, la explotación de sesgos algorítmicos o la delegación cognitiva, entre otros fenómenos que alteran las bases epistémicas del espacio público.

Un punto de partida adecuado para esta tarea lo ofrece el conjunto amplio y creciente de directrices éticas sobre inteligencia artificial. Jobin, Ienca y Vayena (2019) realizaron un análisis sistemático de 84 documentos elaborados por gobiernos, corporaciones y organismos multilaterales, e identificaron una convergencia global en torno a cinco pilares axiológicos. Beneficencia, asociada a la promoción del bienestar. No maleficencia, ligada a la evitación del daño. Autonomía, entendida como respeto por la autodeterminación humana. Justicia, orientada a garantizar equidad. Explicabilidad, destinada a asegurar inteligibilidad de los sistemas. La convergencia resulta alentadora, aunque Brent Mittelstadt (2019) advierte con acierto que los principios, por sí solos, carecen de fuerza suficiente para garantizar una IA ética. Su nivel de abstracción dificulta la traducción operativa y puede derivar en prácticas de *ethics washing*, un "lavado de manos" ético, cuando faltan mecanismos concretos de supervisión y *accountability* sustentados en evidencia verificable.

En consecuencia, un marco de gobernanza capaz de enfrentar la delegación de decisiones morales en sistemas autónomos requiere concebirse como un entramado multinivel. Combina principios generales con estándares técnicos, auditorías algorítmicas, marcos legales sensibles al contexto, junto con una supervisión humana crítica y sostenida. Desde un modelo de responsabilidad distribuida, se vuelven delineables los vínculos y ámbitos de corresponsabilidad entre los actores de la infoesfera. En primer término, diseñadores y desarrolladores, sobre quienes recae la exigencia de incorporar ética desde el diseño, *ethics by design*. Este principio excede la corrección de sesgos en datos (O'Neil, 2016). Supone una reflexión sistemática sobre los valores inscritos en el diseño y en el funcionamiento de sistemas algorítmicos (Verbeek, 2011). Considera de qué modo las arquitecturas técnicas moldean agencia y responsabilidad, a la vez que promueve explicabilidad y evaluación ética antes del despliegue. El propósito

consiste en avanzar hacia sistemas trazables, confiables y normativamente transparentes, en lugar de optimizar de forma ciega métricas de rendimiento tecnocrático (Marcus & Davis, 2020).

A su vez, la responsabilidad individual vinculada al desarrollo y al uso de sistemas de inteligencia artificial se institucionaliza a través de las organizaciones que los diseñan, despliegan y administran. Su tarea consiste en sostener una cultura de desarrollo ético que vaya más allá del cumplimiento formal. Ello implica el fortalecimiento de comités internos de ética con atribuciones decisorias reales, mecanismos de protección para denunciantes (*whistleblowers*) que alerten sobre prácticas riesgosas, además de auditorías externas e independientes aplicadas a los sistemas algorítmicos. En paralelo, la transparencia sobre el uso de datos y sobre el comportamiento efectivo de los modelos adquiere un valor central. Esta exigencia resulta especialmente atingente en modelos de lenguaje a gran escala (LLM), propensos a producir información falsa o alucinaciones psicotécnicas (Mitchell & Krakauer, 2023), puesto que de ella dependen la fiabilidad del sistema y la rendición de cuentas.

El margen de acción de las corporaciones queda enmarcado por la labor de gobiernos y organismos reguladores, llamados a establecer reglas constitutivas que delimiten los alcances de la moralidad delegada. Esa tarea requiere marcos legales robustos de protección de datos (Ienca & Vayena, 2020), prohibiciones explícitas frente a aplicaciones éticamente inadmisibles, como los sistemas de puntuación social o las armas autónomas letales carentes de supervisión humana regulada, junto con agencias de supervisión dotadas de competencia técnica, autonomía institucional y capacidad sancionatoria. Un referente reciente de este enfoque es la Ley de Inteligencia Artificial de la Unión Europea (2024), estructurada en torno a una gestión del riesgo proporcional a la criticidad de cada sistema artificial. La eficacia de este ecosistema normativo e institucional depende, además, de la participación activa de usuarios y sociedad civil. Los primeros requieren alfabetización crítica en torno a la IA, de modo que puedan identificar sesgos, resistir la complacencia moral (Kaas, 2024) y exigir mecanismos de transparencia. La segunda, por medio de la academia, el periodismo y organizaciones cívicas, cumple una función de control público frente a abusos de poder potenciados por la IA (Crawford, 2021), contribuyendo a orientar el desarrollo tecnológico hacia la justicia y la equidad social. En un escenario de gobernanza compleja, adquiere mayor urgencia la formulación de nuevos derechos humanos para la era digital y neurotecnológica. A medida que la IA se integra con tecnologías capaces de registrar e influir en la actividad cerebral, Ienca y Andorno (2017) han propuesto una jurisprudencia destinada a resguardar cuatro derechos fundamentales, libertad cognitiva, privacidad mental, integridad mental y continuidad psicológica. Estos *neurorights* buscan proteger el núcleo de la identidad y la autonomía personal frente a posibles formas de manipulación algorítmica. El debate sobre los llamados “derechos de los robots” (Gunkel, 2018) puede parecer marginal frente a crisis contemporáneas como la guerra, el hambre o la desigualdad proveniente, precisamente, de las grandes tecnológicas. Aun así, la reflexión sobre nuevos derechos humanos orientados a mitigar riesgos asociados a tecnologías disruptivas constituye una tarea ética y política de primer orden.

Por último, cualquier arquitectura de gobernanza éticamente informada ha de preservar capacidad de ajuste ante la evolución constante de la tecnología. La irrupción

de la IA generativa y de los modelos fundacionales ha desplazado las coordenadas del debate, abriendo interrogantes sobre propiedad intelectual, desinformación y equidad (Danaher, 2024). En consecuencia, la gobernanza puede entenderse como un proceso iterativo de tecnorregulación, capaz de aprender, corregir desvíos y reconocer márgenes de error como parte de su dinámica. El horizonte normativo final apunta a consolidar un esquema en el que la innovación tecnológica quede subordinada a criterios de responsabilidad ética y legitimidad pública, integrando ambas dimensiones en una trayectoria de progreso humano sostenido.

7. Conclusiones

El presente artículo siguió un recorrido deductivo que va desde los supuestos ontológicos de la inteligencia artificial hasta las implicaciones prácticas de su gobernanza ética en la era digital. La tesis que orienta la argumentación sostiene que una ética de la IA, teóricamente informada y empíricamente aplicable, exige superar concepciones de agencia apoyadas en conciencia e intencionalidad. Al adoptar una noción informacional en la que agencia es definida por interactividad, autonomía y adaptabilidad dentro de la infoesfera, se establece un fundamento actualizado para enfrentar los desafíos digitales contemporáneos. En ese marco, la IA contemporánea puede comprenderse como una ingeniería altamente sofisticada orientada a reproducir resultados de la conducta inteligente mediante medios distintos a los de una mente consciente. Su prestancia epistemológica y ética reside en la capacidad de desplazar los bordes entre acción, decisión y responsabilidad.

Este encuadre atenúa las derivas antropomórficas y devuelve la reflexión ética a su punto de apoyo sustantivo; los efectos que agentes funcionales producen en el mundo. La ontología informacional propuesta por Floridi ofrece un vocabulario adecuado para describir esta condición emergente, en la que tecnologías de tercer orden median la experiencia humana y transforman el entorno que habitamos, impulsando una reontologización de la existencia, de los fenómenos que allí aparecen, además de las formas de acción que constituyen el ecosistema digital.

Conforme a lo expuesto, el análisis abordó el problema de la brecha de responsabilidad. Así pues, el interés se concentra en comprender la complejidad de la moralidad delegada mediante las nociones de responsabilidad irreprochable y responsabilidad distribuida, en lugar de reducir el problema a la búsqueda de un culpable aislado. Reconocer al agente artificial como fuente causal de una acción permite trazar la cadena de efectos y, a partir de ella, distribuir atribuciones entre las entidades implicadas en diseño, implementación y uso. Este enfoque requiere una condición mínima, explicabilidad técnica, junto con un compromiso vinculante con la transparencia. El argumento incorpora, además, aportes de la psicología moral empírica. Investigaciones sobre expectativas humanas frente a la IA, como el experimento de la Máquina Moral y los estudios sobre la paradoja utilitarista, sugieren que el diseño ético difícilmente puede desarrollarse al margen de las condiciones sociales que lo enmarcan. Intuiciones, sesgos y disposiciones psicológicas forman parte indexical del contexto epistémico y normativo en el que estos sistemas actúan. En tal clave, el objetivo pasa por orientar el desarrollo de sistemas capaces de cooperar virtuosamente con la inteligencia humana, contribuyendo a mitigar sesgos cognitivos, entre ellos la com-

placencia moral, en vez de perseguir la aspiración tecnocrática de fabricar máquinas moralmente perfectas.

En rigor, una ética de la IA se funda en la corresponsabilidad. Dentro de la infoesfera, la acción emerge de una coproducción entre personas, algoritmos e infraestructuras técnicas, lo que vuelve inevitable una distribución efectiva de responsabilidades. En el plano epistémico, los datos sociales describen tendencias y orientan el diseño, aunque la determinación de lo correcto remite, finalmente, a razones públicas y a principios garantistas de justicia y evitación del daño. Pasar del marco a la práctica, exige por ello, cooperación entre gobiernos, empresas, diseñadores y ciudadanía, con el fin de sostener un ecosistema tecnológico compatible con un desarrollo éticamente sostenible. Así las cosas, la inteligencia artificial se presenta como espejo de disposiciones colectivas, refleja sesgos, preferencias, aspiraciones morales de quienes la construyen y utilizan, haciendo visibles tensiones propias de la agencia humana en la transición algorítmica contemporánea.

El *quid* de la cuestión, entonces, concierne al tipo de sociedad que buscamos construir con estas tecnologías. Fundamentar la ética de la inteligencia artificial en una comprensión precisa de la agencia artificial, junto con una distribución equitativa de la responsabilidad, establece condiciones necesarias para asumir esa tarea con prudencia, lucidez crítica y sentido de futuro.

Referencias

Allen, C., Varner, G., & Zinser, J. (2000). *Prolegomena to any future artificial moral agent*. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251–261. https://doi.org/10.1080/09528130050111428

Almendros, L. S., & Echeverría, J. (2019). Ontología y epistemología de la infoesfera. Una interpretación de la filosofía de la información de Luciano Floridi. *Revista de Filosofía*, *44*(2), 263–279. http://hdl.handle.net/20.500.11912/9197

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? En *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bringsjord, S., & Govindarajulu, N. S. (2020). Artificial Intelligence. En E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.). Stanford University. https://plato.stanford.edu/archives/win2020/entries/artificial-intelligence/

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019). Motivations and risks of machine ethics. *Proceedings of the IEEE*, *107*(3), 562–574. https://doi.org/10.1109/JPROC.2018.2865996

Cisek, P. (1999). Beyond the computer metaphor: Behaviour as interaction. *Journal of Consciousness Studies*, *6*(11–12), 125–142.

Coeckelbergh, M. (2020a). *AI ethics*. The MIT Press.

Coeckelbergh, M. (2020b). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Corbí, J., & Prades, J. (1999). *El conexionismo y su impacto en la filosofía de la mente*. Ariel Filosofía.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, *29*(3), 245–268. https://doi.org/10.1007/s13347-015-0211-1

Danaher, J. (2024). Generative AI and the future of equality norms. *Cognition*, *251*, 1–6. https://doi.org/10.1016/j.cognition.2024.105906

Danaher, J., & Sætra, H. S. (2022). Technology and moral change: The transformation of truth and trust. *Ethics and Information Technology*, *24*(35), 1–16. https://doi.org/10.1007/s10676-022-09661-y

Epstein, R. (2016, 18 de mayo). *The empty brain*. Aeon. https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

Floridi, L. (2007). A Look into the Future Impact of ICT on Our Lives. *The Information Society*, *23*(1), 59–64. https://doi.org/10.1080/01972240601059094

Floridi, L. (2013). Technology's in-betweeness. *Philosophy & Technology, 26*(2), 111–115. https://doi.org/10.1007/s13347-013-0106-y

Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 1–13. https://doi.org/10.1098/rsta.2016.0112

Floridi, L. (2024). *Ética de la inteligencia artificial*. Herder Editorial.

Floridi, L. (2025). AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis. *Philosophy & Technology*, *38*, 1–27. https://doi.org/10.1007/s13347-025-00858-9

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Gigerenzer, G. (2022). *How to Stay Smart in a Smart World: Why Human Intelligence Still Beats Algorithms*. MIT Press.

Gunkel, D. J. (2018). *Robot Rights*. The MIT Press.

Hendrycks, D., Song, D., Szegedy, C., Lee, H., Gal, Y., Brynjolfsson, E., Li, S., Zou, A., Levine, L., Han, B., Fu, J., Liu, Z., Shin, J., Lee, K., Mazeika, M., Phan, L., Ingebretsen, G., Khoja, A., Xie, C., ... Bengio, Y. (2025). A definition of AGI. arXiv:2510.18212, 1–57. https://doi.org/10.48550/arXiv.2510.18212

Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy*, *13*(1), 1–27. https://doi.org/10.1186/s40504-017-0050-1

Ienca, M., & Vayena, E. (2020). On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature Medicine*, *26*(4), 463–464. https://doi.org/10.1038/s41591-020-0832-5

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kaas, M. H. L. (2024). The perfect technological storm: moral complacency and the risk of new atrocities. *Ethics and Information Technology*, *26*, 1–12. https://doi.org/10.1007/s10676-024-09788-0

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications*, *157*, 17–24.

Marcus, G., & Davis, E. (2020). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), 1–5. https://doi.org/10.1073/pnas.2215907120

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2). https://doi.org/10.1177/2053951716679679

Myers, S., & Everett, J. A. C. (2025). People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors. *Cognition*, *256*, 1–18. https://doi.org/10.1016/j.cognition.2024.106028

Nyholm, S. (2020). *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Rowman & Littlefield.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Rabossi, E. (1995). *Filosofía de la mente y la ciencia cognitiva*. Alianza Editorial.

Ruiz Osuna, P. (2025). *La personalidad jurídica de los autómatas inteligentes*. Aranzadi La Ley, S.A.U.

Sinnott-Armstrong, W., & Skorburg, J. A. (2021). How AI can aid bioethics. *Journal of Practical Ethics*, *9*(1), 21–40. https://doi.org/10.3998/jpe.1175

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Taddeo, M., & Floridi, L. (2005). Solving the symbol grounding problem: a critical review of fifteen years of research.*Journal of Experimental & Theoretical Artificial Intelligence*, *17*(4), 419–445. https://doi.org/10.1080/09528130500284053

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Unión Europea. (2024). *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo*. Diario Oficial de la Unión Europea. Recuperado de http://data.europa.eu/eli/reg/2024/1689/oj

Varela, F. (1990). *Conocer: Las ciencias cognitivas: tendencias y perspectivas*. Gedisa.

Verbeek, P.-P. (2006). Materializing Morality: Design Ethics and Technological Mediation. *Science, Technology, & Human Values*, *31*(3), 361–380. https://doi.org/10.1177/0162243905285847

Verbeek, P.-P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press.

Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.