ÉTICA DE LA INTELIGENCIA ARTIFICIAL

Rafael Amo Usanos
Sara Lumbreras Sancho
Íñigo Navarro Mendizábal
Directores

ACCESO PREMIUM
Libro + ebook

Colección
IA, Robots, y Bioderecho

Dykinson, S.L.

ÉTICA DE LA INTELIGENCIA ARTIFICIAL

Colección
IA, ROBOTS, Y BIODERECHO

ÉTICA DE LA INTELIGENCIA ARTIFICIAL

Rafael Amo Usanos
Sara Lumbreras Sancho
Íñigo Navarro Mendizábal
(Directores)

Dykinson, S.L.

Este libro ha sido sometido a evaluación por parte de nuestro Consejo Editorial

Para mayor información, véase *www.dykinson.com/quienes_somos*

Editorial DYKINSON, S.L. Meléndez Valdés, 61 - 28015 Madrid
Teléfono (+34) 91 544 28 46 - (+34) 91 544 28 69
e-mail: info@dykinson.com
http://www.dykinson.es
http://www.dykinson.com

ISBN: 979-13-7047-159-0
Depósito Legal: M-9704-2026
DOI: https://doi.org/10.14679/4908

ISBN electrónico: 979-13-7047-255-9

Preimpresión por:
Besing Servicios Gráficos S.L.
e-mail: besingsg@gmail.com

Índice

CAPÍTULO 2.
FUNDAMENTOS FILOSÓFICOS DE LA IA

Rafael Amo Usanos
Universidad Pontificia Comillas

Sara Lumbreras Sancho
Universidad Pontificia Comillas

1. INTRODUCCIÓN

La IA, en cuanto objeto de la filosofía, puede ser abordada por diversas ramas de ésta. La primera y fundamental es la filosofía de la mente, pues es ella la que le aporta el imaginario para crear los conceptos básicos con los que construir una teoría que dé soporte a su intención de emular la mente humana. Esa imitación es la que subyace a la pregunta fundamental de Turing (*¿Pueden las máquinas pensar?)* y a la intención de aquel proyecto de investigación de verano en Dartmouth que le dio nombre a esta disciplina en 1956. La segunda es la epistemología y filosofía de la ciencia, puesto que es fundamental abarcar la cuestión de qué tipo de conocimiento puede obtenerse a través de los métodos, fundamentalmente estadísticos, de la IA. La tercera, la filosofía de la tecnología, ya que la IA es una tecnología y por ello, parafraseando a Melvin Kranzberg, no es *ni buena ni mala, pero tampoco es neutra*. Esta estructura es la que posibilita la entrada en acción de la cuarta rama de la filosofía que se ocupa de la IA, la ética, y que debe tratar no sólo de sus aplicaciones sino también de sus desarrollos. Por último, la antropología, pues los avances de la IA pueden llevar a repensar la naturaleza humana.

2. FILOSOFÍA DE LA MENTE

La filosofía de la mente, perteneciente a la vieja rama de la filosofía natural (Broncano, 1995, p. 11), es el «estudio de los procesos mentales y de los supuestos teóricos subyacentes a las nociones mismas de mente, psique, espíritu, cognición y, en general, de todos los procesos vagamente denominados mentales» (Filosofía de la

mente, s.f.). Es una antigua rama de la filosofía en la que se han producido grandes novedades en este siglo, no sólo como consecuencia del intento más o menos logrado de conseguir máquinas inteligentes, sino como causa de ello. Es decir, que sus avances son los que sustentan a modo de paradigma o imaginario la posibilidad de la IA; aunque también es cierto que, dada la epistemología retroalimentadora de la filosofía natural, la IA ha contribuido al avance de la filosofía de la mente junto con la lingüística, la psicología, las ciencias cognitivas, la robótica, la biología de la evolución, la etología, la antropología, las neurociencias, la neuropsicología, la neurofisiología, la neurocomputación, la lógica y la teoría de la computabilidad, la matemática de los sistemas dinámicos y la filosofía del lenguaje.

La filosofía de la mente aporta al estudio de la IA una reflexión sobre la naturaleza de los procesos cognitivos humanos. Para presentar sus principales contribuciones, en primer lugar se realizará un breve repaso de las taxonomías de la IA, que permitirá centrar la cuestión en su verdadera naturaleza. En segundo lugar, se llevará a cabo un estudio de la filosofía de la mente que soporta la IA realmente existente; y, en tercer lugar, se ofrecerá un balance que permite conocer el alcance filosófico de la IA.

2.1. Taxonomías de la IA

La IA puede ser considerada desde tres puntos de vista, de donde nacen las tres posibles taxonomías que nos interesan en el contexto de la filosofía de la mente.

La primera es la que considera, en atención a sus capacidades, una IA fuerte frente a una IA débil. Así, «de acuerdo con la distinción introducida por John Searle, hay dos modalidades de IA: la IA "fuerte" (frecuentemente confundida con la IA general, porque están relacionadas como veremos) y la IA débil (confundida con la IA específica)» (Madrid, 2024, p. 48). La IA fuerte supone el isomorfismo entre mente y ordenador, frente a la IA débil que «solo admite y explota una cierta analogía entre la mente y el ordenador» (Madrid, 2024, p. 49; Inteligencia artificial, s.f.). Así, la IA débil sólo es funcionalmente equivalente a la humana, es decir, llega a una solución idéntica, pero a través de procesos distintos, que, por supuesto, no necesitan implicar una experiencia consciente. La IA fuerte, por el contrario, sería la construcción de una mente humana artificial que fuera en todo similar a la nuestra, incluyendo la experiencia.

La segunda taxonomía se hace en atención a la extensión de su alcance. Así la IA específica es la que es capaz de imitar tareas específicas de la inteligencia humana (visión artificial, traducción, etc.); frente a la IA general, que sería capaz de tratar cualquier problema que se le presentase. Sería una IA con lo que llamamos a veces *sentido común*, característico de la inteligencia humana (Bonden, 2022, p. 29). Para muchos, una IA general debería ser consciente, y la IA general sería sinónimo entonces de la IA fuerte.

La tercera taxonomía se realiza en atención al modo en el que se procesa la información. Para Boden (2022, p. 15) tenemos cinco categorías: la IA simbólica, la subsimbólica, la conexionista, la evolutiva, la de los autómatas celulares y la de los sistemas dinámicos. De estas categorías, las dos primeras son las más interesantes para nuestro estudio.

2.2. La IA simbólica y subsimbólica

En sus inicios, a principio de los cuarenta, la investigación en IA responde a dos modelos fundamentales, a saber: el simbólico y el subsimbólico. Por un lado, los modelos simbólicos intentan simular las capacidades cognitivas de los seres humanos mediante el procesamiento de fórmulas sintácticas aplicadas a símbolos, pues se entiende que tales capacidades descansan necesariamente en un sistema de representación cuyos elementos básicos reflejan la estructura sintáctica del lenguaje. La subsimbólica es capaz de procesar únicamente vectores numéricos, ¡pero ello nos ha llevado a los mayores logros de la IA hasta este momento!

Por otro lado, la aplicación de estructuras sintácticas a símbolos contrasta con los modelos conexionistas, que encuentran en la estructura sináptica del cerebro la clave para simular la inteligencia humana y, por ello, tratan de imitar la actividad inteligente a partir de la elaboración de redes de conexiones entre unidades muy simples (Corbí-Prades, 1995, 151).

Estos dos modelos han tenido un recorrido desigual. El simbólico fue desarrollado al comienzo de la andadura de la IA, pero no logró los éxitos que de él se esperaban. Por el contrario, el conexionista, basado en las redes neurales, sustenta al modelo actual y ha dado muchos más frutos. A continuación, se describe la filosofía de la mente que los sostiene.

2.2.1. Filosofía de la IA simbólica

La IA simbólica es aquella que intenta emular a la mente humana suponiendo que la inteligencia es fundamentalmente un proceso formal basado en reglas. Basándonos en estas reglas interpretamos los datos que se nos ofrecen y obtenemos respuestas a nuestras preguntas (Cobo-Lloret, 2023, p. 19). Jerry Fodor y su teoría representacional (que sostenía que el pensamiento era una forma de manejo de símbolos) y la teoría modular de la mente es uno de los inspiradores de esta forma de entender la IA; y Hubert Dreyfus, uno de sus mayores críticos, al explicar que la mente humana era mucho más que eso.

La IA simbólica tiene un triple fundamento en las ciencias cognitivas. El primero en la psicología cognitiva que postula la existencia de los estados mentales, frente

a la psicología conductista. El segundo, en el funcionalismo analítico que «intenta comprender cómo se articulan entre sí los diversos estados mentales de un sistema» (Amo, 2007, p. 150), y que, sin negar el contenido del estado mental, no le presta atención. El tercero en la filosofía del lenguaje de corte positivista y analítico que suponen que el contenido del lenguaje (la articulación de estados mentales) es reductible a la lógica formal. Así, «los modelos simbólicos realizan una aportación clave en este punto, pues muestran cómo la semántica puede afectar al mundo a través de la sintáctica. Es decir, el único modo de explicar la eficacia causal de los contenidos mentales es postular que éstos se hallan codificados en una estructura sintáctica» (Corbí-Prades, 1995, 157).

La crítica a la IA simbólica la elaboró Dreyfus, en cuatro niveles: i.) el biológico, porque supone que «el cerebro es una máquina de estado discreto equivalente a una computadora electrónica» (Carabantes, 2016, p. 280); ii.) el psicológico, porque «apoyándose en una imagen dualista de la mente, establece que esta puede ser considerada como un dispositivo que opera con bits de información de acuerdo a reglas formales, igual que un programa informático» (Carabantes, 2016, p. 280); iii.) el epistemológico, ya «que da por sentado que todo el conocimiento puede ser formalizado y, por tanto, computado por un ordenador» (Carabantes, 2016, p. 280); y, el iv.) el ontológico, porque supone que «la realidad consiste en un conjunto de hechos independientes entre sí desde un punto de vista lógico» (Carabantes, 2016, p. 280).

2.2.2. Filosofía de la IA subsimbólica

La IA subsimbólica es aquella que intenta emular la mente humana suponiendo que esta es el resultado de las conexiones neuronales. Piensa que en el cerebro –al que debe imitar la IA– cada neurona es la responsable de aportar un rasgo de la realidad (unidad) al contenido del estado mental que se está representado; y que ese estado mental es el fruto de la interconexión entre las unidades aportadas por cada neurona. Esto implica suponer que la realidad puede ser descrita al ser descompuesta en unidades y sus interconexiones (Cobo-Lloret, 2023, pp. 17-27). Entre los autores que la sustentan se cuentan W. Pitts, W. McCulloch, M. Minsky y S. Papert; y entre sus críticos, el citado Jerry Fodor (Inteligencia artificial, s.f.).

Enmarcada en el funcionalismo analítico, la IA subsimbólica se orienta principalmente a comprender la articulación de los estados mentales, sin prestar especial atención a sus contenidos. La diferencia con la IA simbólica radica, al menos, en dos aspectos fundamentales: su posición respecto al contenido de los estados mentales y su concepción de la filosofía del lenguaje.

Sobre la primera, la IA subsimbólica opta por el eliminativismo. Con esta opción genera la idea conexionista de la mente para la cual «los contendidos mentales no se

codifican ya en fórmulas sintácticas, sino en redes de actividad» (Corbí-Prades, 1995, p. 162). Sobre la segunda, afirma que la realidad no se comprime en proposiciones, sino en unidades y sus interconexiones, en la que «cada unidad representa un rasgo del mundo» (Corbí-Prades, 1995, p. 160) y el contenido de una proposición se encuentra en la red de unidades.

Este modelo ha dado los grandes resultados que estamos viendo. Según Cobo y Llorente «la clave del éxito de una red neuronal radica en los cálculos que se realizan entre cada neurona y la forma en que se ajustan los pesos de las conexiones entre ellas. Estos pesos se ajustan de manera que el modelo pueda aprender patrones y correlaciones en los datos, lo que permita hacer predicciones más precisas» (Cobo-Llorente, 2023, 24).

2.3. Balance filosófico de la IA

Se trata ahora, no de considerar las bases filosóficas que sustentan a la IA, sino de considerar las consecuencias que la IA tiene para conceptos filosóficos de primer orden como inteligencia o ser humano.

2.3.1. La IA y el concepto de inteligencia

Tanto el modelo simbólico como el subsimbólico renuncian a la cuestión del contenido de los estados mentales y, con ello, a conocer el contenido del lenguaje que se convierte en algo puramente formal, quizá porque, como afirman Bertolaso y Marcos, parten de la hipótesis de la isomorfía entre *physis* y *logos*: «lo que está en el fondo es la convicción de que nuestro pensamiento y nuestro lenguaje, cuando son verdaderos (en teoría) o funcionales (en la práctica), se corresponden de algún modo con la forma del mundo» (Bertolaso-Marcos, 2024, p. 20).

Entonces para la IA basta con manejar la sintaxis: las diferencias entre un modelo y otro estriban en el modo de conocer las reglas sintácticas. Uno, el simbólico, piensa que estas reglas se conocen previamente y se deben programar para que la IA haga su trabajo. Otro, el subsimbólico, parte de la convicción de que se pueden inducir, esto es, inferir buscando patrones en las unidades menores.

Estos presupuestos de los modelos de IA suponen la renuncia inicial a los otros dos niveles de lenguaje (semántica y pragmática). Esta renuncia, en ambos casos, procede del funcionalismo analítico –verdadero suelo filosófico de la IA al que no le importa el contenido del estado mental–, y arroja, al menos, dos importantes consecuencias sobre el concepto de inteligencia:

i. La primacía de la sintaxis

La posición común de los dos modelos de IA en lo que respecta al formalismo del lenguaje y a su, consiguiente, aceptación del papel único o preponderante de la sintaxis en la inteligencia relegando a la semántica o la pragmática, supone para algunos autores un límite que impide alcanzar al IA fuerte y general, que es el sueño último de esta tecnología.

Esta es la famoso crítica de Searle a los modelos de IA, el argumento de la habitación china. La reducción del lenguaje a lógica formal obvia los significados (sea cual sea el origen de ellos): «Hoy en día los ordenadores realizan una manipulación sintáctica (M1) sin componentes semánticos (M3 por mediación de M2). En palabras de Searle: 'Las mentes son más que sintácticas. Las mentes son semánticas'» (Madrid, 2024, p. 58).

ii. El problema del aprendizaje

Ya se ha dicho que la IA simbólica opta por suponer que la sintaxis es la estructura de la mente y desde ahí se deduce el conocimiento. Puede que esta sea la razón de sus pocos éxitos.

Por su parte, la IA subsimbólica supone que la sintaxis se puede inducir del análisis de los patrones estadísticos en miles de billones de datos. Esto le incapacita para aprender en el sentido amplio de la palabra porque el aprendizaje no se hace solo por inducción, ni siquiera por deducción (que se lo digan a los científicos medievales), sino que, en palabras de Larson, que toma prestadas del semiólogo Pierce, también por abducción:

> La abdución es un razonamiento apogógico (por decirlo en términos aristotélicos), esto es, una clase de razonamiento –distinto de la deducción y de la inducción– por el que hacemos conjeturas ante un hecho sorprendente en función del contexto y la experiencia tomando partido por una de ellas. Mediante la abducción no sólo podemos inferir la mejor explicación de un fenómeno, sino que también podemos realizar inferencias causales, seleccionando entre causas o factores enfrentados en la determinación de un efecto (Madrid, 2024, pp. 82-83).

Esta forma de conocer, que subraya la informalidad y la contextualidad del conocimiento, como pusieron de manifiesto Hubert y Dreyfus, lo que pone sobre la mesa es el papel de los conocimientos implícitos (cosmovisión, cultura) que se adquieren por endoculturación y que tanta importancia tienen en el día a día de las personas.

2.3.2. La IA y la antropología

A una conclusión similar a la anterior, es decir que los sistemas de IA no pueden aprender ni conocer como los humanos, se puede llegar desde la antropología. El aprendizaje humano, afirma Padial, puede ser por medio de la razón o por medio de hábitos y costumbres, ahora bien, para que eso ocurra, «que las máquinas sean susceptibles de hábitos requeriría que tuviesen un cuerpo del que tomar posesión, y un inconsciente que mediara entre ese cuerpo y una autoconciencia intelectual o reflexiva» (Padial, 2019, 200).

Por otra parte, la opción por el funcionalismo analítico frente al teleológico como suelo filosófico para los modelos de IA, tiene sus consecuencias. La primera es que la renuncia de los modelos simbólicos y subsimbólicos a considerar el contenido del estado mental y la semántica y pragmática del lenguaje, supone dejar fuera de la definición de inteligencia (humana y artificial) las nociones de emoción y de *qualia* (experiencia cualitativa, como por ejemplo la experiencia de un color, que no es equivalente al conocimiento del código RGB que lo representa o de la música, que no es equivalente a los datos en cualquier formato que lo codifiquen). La segunda es que la opción por el funcionalismo arrastra la metáfora software y hardware con el consecuente dualismo antropológico que devalúa el cuerpo dando a la inteligencia, incluso en soporte informático, la primacía en lo humano.

Así, la imagen de lo humano que se desprende del supuesto isomorfismo entre IA fuerte y mente humana es que el ser humano está dividido en dos elementos al modo cartesiano, y que la emoción presenta un difícil encaje en la comprensión del ser humano. A esta cuestión se volverá al final de este capítulo.

3. EPISTEMOLOGÍA Y TEORÍA DE LA CIENCIA

Los avances en la IA han provocado un cambio como el que se produjo con el nacimiento de la *Nuova sciencia* y que culminó Francis Bacon en el *Novum Organum*. Entonces se pasó de la deducción a la inducción como método científico. Ahora, esta inducción evoluciona a la observación masiva de datos y la extracción de patrones a partir de ellos.

Es clave darnos cuenta de que los modelos de aprendizaje automático no contienen creencias proposicionales, sino que simplemente representan funciones estadísticas entrenadas para maximizar rendimiento predictivo dentro de un dominio dado, es decir: no buscan el descubrimiento de la verdad sino la maximización de un uso dado. Su estatuto epistemológico es, por tanto, instrumental y además, su utilidad está condicionada a la calidad de los datos, los protocolos de verificación y la supervisión humana que los contextualiza (Bender y Koller, 2020).

La inducción está presente en la mayoría de los desarrollos de la IA que acaparan la atención actualmente, y que pertenecen a la IA subsimbólica: redes neuronales profundas, árboles de decisión o métodos de vectores soporte se especializan en generalizar regularidades a partir de ejemplos.

Dentro de la IA simbólica tenemos la programación lógica inductiva –con obras clásicas como FOIL (*First Order Inductive Learner*)– que muestra que incluso pueden inducirse reglas de primer orden simplemente a partir de ejemplos (Quinlan, 1990). Además, cuando el conocimiento previo está formalizado en reglas, la IA es capaz de realizar deducciones. Este fue el origen de buena parte de la GOFAI que describíamos arriba. Nos encontramos con motores de inferencia como Prolog (desarrollado en los años ochenta pero aún en uso), o los demostradores de teoremas, que derivan consecuencias lógicamente necesarias de unas premisas dadas, garantizando validez siempre que las premisas sean correctas (de Moura y Bjørner, 2008).

Es necesario subrayar que, en la inducción realizada en la IA subsimbólica, se extraen únicamente patrones estadísticos que, por su propia naturaleza, no son capaces de descubrir relaciones causales. Sin embargo, en los años recientes ha venido desarrollándose el campo revolucionario de la inferencia causal, que ha introducido herramientas específicas para identificar o descartar estas relaciones. Las redes bayesianas dirigidas, en los que se modelan flechas causales, pueden estimar los efectos de una hipotética intervención. El *do-calculus* de Judea Pearl consiste en tres reglas matemáticas que permiten reescribir expresiones con el operador *do* (que para él representa hacer una intervención, que se contrapone a observar una variable). Usándolo, es posible en ocasiones aislar efectos causales únicamente a partir de datos observacionales (Pearl, 2009).

Cuando ese grafo causal es desconocido, algoritmos de descubrimiento causal como PC o GES reconstruyen la estructura mediante independencias condicionales extraídas de los datos (Spirtes, Glymour y Scheines, 2000). Con todo, la validez de este estudio causal descansa en la corrección de los supuestos estructurales: una orientación errónea del grafo o la omisión de variables ocultas puede inducir sesgos graves que el formalismo no es capaz de corregir. Aun así, las posibilidades que ofrecen estas técnicas, con poco más de una década, hacen que la IA haya pasado de ser un descubridor de correlaciones a una herramienta mucho más profunda en la extracción de conocimiento.

Por un lado, otros rasgos de la IA deben tenerse en cuenta a la hora de analizar su estado epistemológico. Para empezar, su carácter instrumental y, en el caso de la IA subsimbólica, no proposicional implica que la *verdad* se mide pragmáticamente mediante métricas matemáticas de error que el desarrollador escoge (como por ejemplo F1 y otras fórmulas matemáticas). Esto hace que no tenga criterios filosóficos sino técnico-estadísticos.

Por otro lado, la dependencia de los datos hace que aparezcan los sesgos, amenazando no sólo la fiabilidad epistémica (es decir, la garantía de la verdad), sino también la justicia.

Además, la fiabilidad de los modelos es radicalmente contextual puesto que depende de los datos: en dominios cerrados muy bien definidos, como la clasificación radiológica, un sistema puede alcanzar o superar la precisión de expertos. Sin embargo, fuera de ese entorno pierde esta robustez y se hacen imprescindibles protocolos de validación externos. El tema de la fiabilidad ha pasado a ser un problema clave en el caso de los LLM, puesto que las alucinaciones (salidas plausibles pero falsas) son un riesgo al que, por el momento, no se ha conseguido dar una respuesta definitiva.

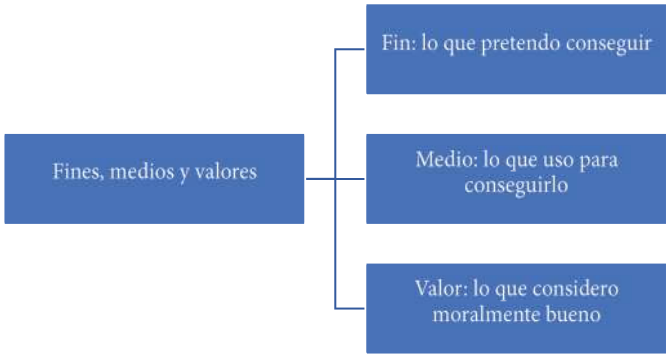
4. FILOSOFÍA DE LA TECNOLOGÍA Y LA IA

La tercera rama de la filosofía que debe hacerse cargo de la reflexión sobre la IA es la filosofía de la ciencia y de la tecnología, ya que fundamentalmente la IA es una tecnología, o, según algunos, una ciencia.

Lo más significativo de la reflexión de la filosofía de la tecnología sobre la IA es la afirmación sobre su carácter ético. Así afirma el Documento vaticano Antiqua et Nova: «La actividad técnico-científica no tiene un carácter neutro, siendo una empresa humana que pone en cuestión las dimensiones humanísticas y culturales del ingenio humano» (Dicasterio ..., 2025, nº 36). Y es que la filosofía de la tecnología hace tiempo que abandonó la tesis de la neutralidad axiológica de la filosofía (Bertolaso-Marcos, 2024, pp. 47-50).

Esto se debe a que la ciencia y la tecnología son considerados como sistemas construidos por agentes que deliberadamente buscan ciertos fines, en función de determinados intereses, para lo cual ponen en juego creencias, conocimientos, valores y normas. Los intereses, los fines, los valores y las normas forman parte también de esos sistemas y son susceptibles de una evaluación moral. Estos fines y valores no son posteriores al resultado científico ni tecnológico, sino que acompañan todo el proceso desde la intención inicial al diseño y realización.

Figura 1:
La tecnología como sistema intencional



Fuente: elaboración propia

Esta forma de comprender los sistemas tecnológicos es la que abre la puerta a la reflexión sobre la ética de la IA (Olivé, 2003, pp 181-223; Olivé, 2001, pp. 45-59). Sin la consideración de los fines y valores, un sistema tecnológico solo se puede medir por su eficacia o ineficacia, si los medios consiguen que el sistema tecnológico logre aquello para lo que se diseñó.

Una derivada posterior de la reflexión sobre la IA desde la filosofía de la tecnología debe abordar el posible papel de la IA en la cultura tecnocrática. Así lo afirma el Documento Antiqua et Nova de los dicasterios vaticanos: «existe el riesgo de que la IA se utilice para promover lo que el Papa Francisco ha llamado "paradigma tecnocrático", que tiende a resolver todos los problemas del mundo sólo con medios tecnológicos. Según este paradigma, la dignidad humana y la fraternidad, a menudo, se dejan de lado en nombre de la eficacia, "como si la realidad, el bien y la verdad brotaran espontáneamente del mismo poder tecnológico y económico"» (Dicasterio para la Doctrina de la Fe & Dicasterio para la Cultura y la Educación, 2025, nº 54).

Este riesgo se puede hacer realidad cuando la eficacia, que es resultado de unos medios adecuados, es el único elemento evaluado en un sistema tecnológico y se olvidan la dignidad y el bien común, que son valores a los que deberían señalar los fines de la tecnología.

5. LA ÉTICA DE LA IA

Como se ha indicado, la IA es un sistema intencional y, por tanto, susceptible de ser evaluado éticamente. La ética no puede dejar pasar el impacto que la IA está te-

niendo en la vida de las sociedades y las personas. La preocupación social por una IA confiable, responsable, es manifiesta, y la propia existencia de este libro lo demuestra. Son múltiples las declaraciones institucionales y las publicaciones científicas. Muchas personas creen que el poder de esta tecnología es tal que, con malos fines puede provocar grandes daños. Otros, por el contrario, piensan que una regulación ética supondrá un límite a sus capacidades. En cualquier caso, se hace necesario que la ética se haga cargo de las posibilidades de la IA.

En un primer momento, se considerará el estatuto epistemológico de la ética de la IA y, en un segundo momento, se ofrecerá un panorama de los nudos éticos gordianos a los que se enfrenta una aplicación de IA en un campo cualquiera.

5.1. Estatuto epistemológico de la ética de la IA

Con esta rotundidad explica este documento de la Comisión europea la naturaleza epistemológica de la ética de la IA: «la ética de la IA es un subcampo de la ética aplicada que estudia los problemas éticos que plantea el desarrollo, despliegue y utilización de la IA. Su preocupación fundamental es identificar de qué modo puede la IA mejorar o despertar inquietudes para la vida de las personas, ya sea en términos de calidad de vida o de la autonomía y la libertad humanas necesarias para una sociedad democrática» (Grupo de Expertos de Alto Nivel sobre la Inteligencia Artificial, 2019, p. 32).

Como afirma Beorlegui, por su naturaleza de ética aplicada «no se trata de partir de una serie de principios generales básicos, para aplicarlos después a diversos campos concretos de la vida humana» (Beorlegui, 2023, p. 19), sino que es necesaria una epistemología propia, como la hermenéutica crítica que propone Adela Cortina.

Esta propuesta consta de dos momentos. El primero es el deontológico en el que se propone un imperativo: «obra de tal modo que tu acción vaya encaminada a sentar las bases (en la medida de lo posible) de una comunidad ideal de comunicación» (Beorlegui, 2023, p. 19). El segundo es el aristotélico en el que el imperativo es modulado en cada campo de realidad específico. De aquí la característica interdisciplinar de la ética aplicada. Esta modulación consiste en «descubrir en cada campo de la ética aplicada las máximas y valores exigidos en ese ámbito» (Beorlegui, 2023, p. 20).

De la abundante bibliografía sobre el tema se puede deducir que el armazón de esta epistemología de la ética de la IA está compuesto por valores, que son bienes a promocionar y proteger por medio de principios, que son los «imperativos éticos que los profesionales de la IA deben esforzarse en todo momento por observar» (Grupo de Expertos de Alto Nivel sobre la Inteligencia Artificial, 2019, p. 47). Estos principios se hacen operativos por medio de requisitos, las condiciones que la IA debe cumplir en su desarrollo y uso para que se puedan poner por obra los principios que promocionan y protegen los valores. Y junto a todo ello, las virtudes, que son las dis-

posiciones morales de quien utiliza la tecnología para que pueda poner en acto los principios

i. Los valores

En el campo de la axiología el concepto de valor tiene muchas definiciones. En general se puede decir que valor es «todo aquello que debe ser objeto de preferencia o de elección» (Abbagnano, 2004, p. 1071). En el campo de los valores éticos, por encima de teorías particulares, se puede considerar que un valor moral es el bien.

En las declaraciones institucionales y en la bibliografía sobre la ética de la IA se habla de los valores como dignidad, libertad, igualdad, solidaridad, verdad y bien integral.

ii. Los principios

Se entiende por principios aquel marco «de normas generales, derivadas de la moral común, que constituyen un punto de partida adecuado para la reflexión sobre los problemas morales en ética» (Beauchamp-Childress, 2024, p. 79).

Hay un acuerdo bastante generalizado en proponer los cuatro principios clásicos de la bioética como aplicables a la ética de la IA. Así afirma Adela Cortina:

> los principios clásicos serían el de beneficencia, que exigiría ahora poner los progresos al servicio de todos los seres humanos y la sostenibilidad del planeta; el de no-maleficencia, que ordenaría evitar los daños posibles, protegiendo a las personas en cuestiones de privacidad, mal uso de los datos, en la posible sumisión a decisiones tomadas por máquinas y no supervisadas por seres humanos; pero también el principio de autonomía de las personas, que puede fortalecerse con el uso de sistemas inteligentes, y en cuyas manos deben ponerse tanto el control como las decisiones significativas; y, por supuesto, el principio de justicia, que exige distribuir equitativamente los beneficios (Cortina, 2024, p. 69).

Ahora bien, Cortina, añade un quinto principio «la explicabilidad y la rendición de cuentas, y la trazabilidad […] porque los afectados por el mundo digital tiene que poder comprenderlo» (Cortina, 2024, p. 69). De forma similar se expresa Floridi, quien llega a la misma conclusión después de estudiar una gran bibliografía y declaraciones institucionales (Floridi, 2024, pp. 139-158). Sin embargo, al revisar los documentos y la bibliografía parece que ese quinto principio es más bien un requisito técnico, al que sumar otros muchos, que hace posible que los principios protejan y promocionen los valores.

De este modo los cuatro principios de la Bioética, la ética aplicada más consolidada, pueden ser aplicados a la ética de la IA: autonomía, beneficencia, no-maleficencia y justicia.

iii. Los requisitos

Los requisitos, es decir, aquellas condiciones técnicas que debe cumplir los sistemas de IA para que se puedan hacer realidad los principios éticos, con todo lo que esto implica, es el elemento más discutido en la bibliografía. No hay consenso en cuales sean estos requisitos y si alguno de ellos, como se ha comprobado, puede ser considerado un principio ético.

Se puede considerar que estos requisitos son cinco: explicabilidad, autonomía informacional y privacidad, robustez, equidad y sostenibilidad.

En primer lugar, la *explicabilidad* es el más importante de estos requisitos, e incluso para algunos cobra estatus de principio ético. Tiene que ver con todo aquello que hace referencia a la necesidad de que los algoritmos den razón de cómo han llegado a las conclusiones, ya que este camino presenta, en muchos casos dificultades ya que, como se ha indicado, no utilizan una lógica causal, entendida esta en su sentido clásico.

La complejidad de este requisito se comprueba por el enorme campo semántico que utiliza, por la gran cantidad de conceptos que, sin ser idénticos, se refieren a la explicabilidad, aportando matices importantes: transparencia, la trazabilidad, la comunicación, la interpretabilidad, el conocimiento, la inteligibilidad y la auditabilidad.

En segundo lugar, la *autonomía informacional y privacidad*, como las denomina Floridi (2024, p. 231). En este requisito se da cita todo aquello que se refiere a las limitaciones e influencias que los sistemas de IA pueden ejercer sobre la autonomía humana y su privacidad. Este requisito implica la supervisión humana, protección de datos, seguridad y protección, protección de la intimidad, calidad e integridad de los datos.

En tercer lugar, la *robustez*. Este requisito tiene que ver con todo lo referente al daño que los sistemas de IA pueden llegar a causar en los usuarios, por errores de programación o por programación maligna. La robustez implica la seguridad de los algoritmos, la solidez, la precisión, la fiabilidad, la rendición de cuentas o la auditabilidad, entre otros elementos.

En cuarto lugar, la *equidad* que guarda relación con la salvaguarda de la justicia en los resultados de las aplicaciones de IA. Es decir, que estas aplicaciones han de evitar la discriminación y lograr un trato igualitario a las personas. La equidad implica los esfuerzos para evitar los sesgos.

En quinto, y último lugar, el requisito de la *sostenibilidad*. Este requisito implica que los sistemas de IA deben tener en cuenta su impacto en las personas, sociedades y medio ambiente. Así deben diseñarse teniendo en cuenta el respeto y promoción de

la salud física y mental de las personas; sus repercusiones sobre las instituciones, la democracia y la sociedad; y el respeto por el medio ambiente.

Figura 2.
Los requisitos técnicos

Explicabilidad	Referente a las consecuencias de la lógica no causal de la IA
Autonomía y privacidad	Referente al respeto a la autonomía y privacidad por parte de los sistemas de IA
Robustez	Referente al daño que pueden causar las aplicaciones de IA
Equidad	Referente a la necesidad de la no discriminación en las aplicaciones de IA
Sostenibilidad	Referente al impacto de la IA en las personas, sociedades y medio ambiente

Fuente: elaboración propia

iv. Las virtudes

La ética aplicada no puede olvidar a quién la aplica, al sujeto moral, por eso en el esquema de la epistemología de la ética de la IA se deben considerar las virtudes que «son una disposición activa, voluntaria y persistente para practicar el bien en cualquier circunstancia y no de modo ocasional, pudiendo ser perfeccionada a través de la enseñanza y la práctica» (Cruz, 2016, p. 176).

En una ética aplicada, como la ética de la IA, hay que considerar el replanteamiento que de las virtudes hace MacIntyre, para quien las virtudes son «las cualidades que permiten alcanzar los bienes internos o intrínsecos a una determinada práctica» (Cruz, 2016, p. 181). Así las virtudes están en clara conexión con los valores de la IA.

5.2. Panorama de las problemáticas éticas de la IA

El uso de la IA se enfrenta, fundamentalmente a cuatro tipos de problemáticas éticas: los problemas de la algorética; la ética de los usos de la IA en cuanto sistema de recomendación o sistema de decisión; la ética del usuario de la IA; y cómo debe ser tratada la IA.

5.2.1. La algorética

Por algorética se puede entender «las cuestiones éticas que plantean los algoritmos, entendiendo estos como construcciones matemáticas, por su implementación como programas y configuraciones (aplicaciones)» (Floridi, 2024, p. 204). Y es que, como bien afirma Floridi, «los algoritmos no son éticamente neutrales» (Floridi, 2024, p. 205), es decir, son sistemas intencionales y, por tanto, objeto de la ética.

En este campo de la ética de la IA es donde se hacen operativos los cinco requisitos de los que se ha hablado. La bibliografía recoge, al menos, tres problemáticas que se pueden agrupar en lo que se denomina algorética: el problema de la gestión de datos, el de los daños y el de los sesgos (Coeckelbergh, 2021, pp. 75-122).

i. La gestión de datos

Los algoritmos se alimentan de datos y los gestionan para dar un resultado. Este funcionamiento puede generar dificultades éticas en la obtención de los datos, en el hecho de que exista consentimiento y el tipo de consentimiento que sea, es decir, todo lo referente a la privacidad. No se puede olvidar el riesgo de que los humanos puedan ser valorados no por sí mismos, sino como medio para obtener datos.

Ahora bien, el output que genera un algoritmo puede ser utilizado para limitar la autonomía de las personas, dirigiéndolas de forma cuasi inconsciente a una determinada dirección, especialmente si son personas vulnerables.

El requisito algorético de la autonomía y privacidad, junto con el de sostenibilidad, es imprescindible para que el desarrollo y la implementación del algoritmo evite estos abusos.

ii. El problema del daño

Ya desde la Ética a Nicómaco de Aristóteles se habla de la responsabilidad de las acciones: quien actúa debe responder de los resultados de su acción. Ahora bien, «si a una IA se le otorga una mayor capacidad de actuación y asume tareas que solían llevar a cabo los humanos, ¿hemos de atribuirle responsabilidad moral? ¿Quién es responsable de los perjuicios y los beneficios que causa la tecnología cuando los humanos delegan capacidad de actuación y toma de decisiones en la IA? Reformulándolo en términos de riesgo: ¿quién es el responsable cuando algo sale mal?» (Coeckelbergh, 2021, p. 95).

Es aquí donde se hacen operativos los requisitos de explicabilidad y robustez con los que se deben desarrollar los algoritmos, que, junto con la sostenibilidad, debe afrontar la cuestión del daño que puede causar la IA.

iii. El problema de los sesgos

El tercero de los problemas de la algorética que más preocupa es el del sesgo: «cuando una IA toma (o, de forma más precisa, recomienda) decisiones, puede hacerlo de forma desigual, pero puede surgir el sesgo: las decisiones puntuales pueden ser injustas o poco equitativas para individuos o grupos particulares […] El problema del sesgo está a menudo relacionado con las aplicaciones de aprendizaje automático. Y, aunque los problemas de sesgo y discriminación siempre han estado presentes en la sociedad, la preocupación es que la IA pueda perpetuar estos problemas y acrecentar su impacto» (Coeckelbergh, 2021, p. 107).

Los sesgos pueden surgir en todas las fases del algoritmo, el diseño, la prueba y la aplicación y es el requisito técnico de la equidad el que ha de velar por el control de los sesgos.

5.2.2. Los usos de la IA

La IA tiene, fundamentalmente, dos grandes usos, como sistema de recomendación y como sistema de decisión, lo que permite clasificarlos en tres grupos en atención a la relación con la decisión humana: *human-in-the-loop*, *human-on-the-loop*, *human-out-the-loop*.

i. La IA como sistema de recomendación (*human-in-the-loop*)

Los sistemas de IA con supervisión humana (*human-in-the-loop*) son aquellos en los que el ser humano interviene activamente en el proceso de toma de decisiones. La IA asiste, pero no actúa de manera autónoma. Estos sistemas actúan como sistema de recomendación. Un ejemplo puede ser los de ayuda al diagnóstico médico.

Cuando la IA es utilizada por el usuario humano para ayudarle en su decisión que es tomada directamente por el humano, o bien, la decisión cuenta con la supervisión humana, se presentan, al menos, tres tipos de problemáticas: el desarrollo y alimentación de los sistemas de recomendación (aquí nos referimos al problema tratado de los sesgos y el manejo de datos); el uso del sistema de recomendación: la ética del usuario; y la interposición de la tecnología entre los humanos[1].

La tecnología es parte de nuestra vida, incluso para muchos autores forma parte esencial de la cultura, nos hominizó y nos humaniza. Ahora bien, ¿hasta qué punto estamos dispuestos a que se interponga entre dos humanos cuando uno decide sobre

1 La primera de las cuestiones ya se ha tratado al hablar de la algorética. La segunda, se tratará al abordar la cuestión de la ética el usuario.

otro? Es cierto que esta pregunta se puede aplicar a cualquier tecnología que medie entre dos humanos, pero el carácter disruptivo de la IA obliga a planteársela de nuevo.

El problema se hace acuciante en los casos en los que el sistema de IA puede ser confundido con un humano, por ejemplo la IA de cuidado y de compañía, que son asemejables a este tipo de sistemas de recomendación porque en ellos se oscurece el valor de la relación interpersonal que se basa en la empatía y la compasión, y puede dar lugar al posible engaño, hacernos creer que la probabilidad (base de la IA subsimbólica), completamente carente de experiencia subjetiva o de identidad, nos acompaña y alivia la soledad.

Lo dicho hasta aquí, puede ser aplicado también a los sistemas con supervisión humana pasiva, es decir, *human-on-the-loop*. Aquellos en los que la IA toma decisiones de forma autónoma, pero un humano supervisa el proceso y puede intervenir o detenerlo si es necesario.

ii. La IA como sistema de decisión autónomo (*human-out-the-loop*)

Los problemas éticos del uso de la IA se multiplican cuando esta es utilizada como sistema de decisión, es decir, cuando el sistema de IA es el que decide la acción a ejecutar y la ejecuta sin supervisión humana o sin mediación humana.

Fundamentalmente se pueden plantear dos cuestiones: estos sistemas autónomos ¿pueden ser considerados agentes morales?, es decir, pueden tomar decisiones ética; y si es así ¿con qué ética se deberían programar?

A la primera cuestión, se responde fácilmente que no. No pueden ser agentes morales porque no son fuente originante de sus acciones ni pueden valorar. No debemos olvidar que, pese a que la IA pueda resultar funcionalmente equivalente a una persona en determinadas actividades, como el uso del lenguaje, esta equivalencia funcional está lejos de ser una equivalencia ontológica. Como explica Lumbreras (2019), tanto los procesos de elección, como el uso del lenguaje o potencialmente otras capacidades, como la empatía que pueden desarrollar las máquinas no corresponden a una activación auténtica de estas capacidades, porque en la máquina no existe un sujeto que las experimente, no hay una interioridad que acoja estos estados mentales y emocionales. Esto quiere decir que no puede haber tampoco comprensión, creencia, emoción, ni, por tanto, empatía, amor, libertad o responsabilidad.

La IA, en sentido estricto, no actúa por una causa interior, sino que simplemente tiene la capacidad de interactuar con el entorno adaptando la propia conducta sin estar al albur de cualquier estímulo. Esto se podría considerar autonomía en sentido funcional, no ontológico, y se debería llamar independencia externa.

Además, la IA puede elegir, pero no decidir. Dice Antiqua et nova (Dicasterio ..., 2025, nº 46): «El ser humano, en cambio, no sólo elige, sino que en su corazón es capaz de decidir». Elegir supone «un sistema inteligente puede deliberar sobre los planes y ejecutarlos, contando con las informaciones que se le han proporcionado, tanto de hechos como de finalidades»; sin embargo, para decidir se exige que el agente moral esté «dotado de creencias, preferencias, emociones morales (empatía, simpatía, culpabilidad, vergüenza), voluntad libre y la capacidad de reflexionar sobre sus razones morales, modificarlas, justificar las bases normativas de sus decisiones morales y tener en cuenta otros cursos de acción».

Ni siquiera en el supuesto de que se lograse la IA fuerte y general, la IA no decidiría, porque, como pone de manifiesto Zubiri con su comprensión de la inteligencia sentiente, para poder hacerlo hay que valorar y en el proceso de valoración debe formar parte el sentimiento y la voluntad (Amo, 2023, pp. 84-94).

La disciplina conocida como Ética de las máquinas o *Machine Ethics*, es la que se centra precisamente en dotar a los dispositivos autónomos de procedimientos internos que garanticen decisiones moralmente aceptables (Veruggio et al., 2016; Floridi, 2005). Muchas éticas serían posibles. Para presentarlas seguiremos el esquema crítico propuesto en Lumbreras (2017), que estudia todas las posibilidades de introducción de ética en las máquinas a partir de una matriz que las clasifica entre éticas negativas/positivas, y *top-down/bottom-up*. La ética negativa estaría orientada a impedir el daño, y la ética positiva, orientada a maximizar el bien (Handelsman et al., 2009). La distinción entre la orientación de los enfoques contrapone el *top-down* como reglas o fines impuestos externamente, con el *bottom-up* referiría valores o principios que emergerían de la experiencia del agente (Allen et al., 2005; Campbell et al., 2002).

La deontología kantiana ilustra el cuadrante negativo *top-down*: los deberes universales pueden codificarse como reglas, aunque en la práctica surgen inevitablemente conflictos normativos (¿debo mentir para evitar que maten a alguien?) que obligan a jerarquías o meta-prioridades (Kant, 2002), tal como anticipa la literatura de ficción de Asimov (1950).

El consecuencialismo utilitarista representa la ética positiva: su viabilidad depende de traducir la noción de bienestar a una función objetivo-cuantificable (Jackson, 1991). Esta posibilidad parece, de entrada, la más fácil de implementar, pero, en cuanto abandonamos situaciones sencillas, nos damos cuenta de que expresar la bondad de una situación como una función matemática objetiva nos resulta imposible sin reducir la realidad a su descripción más pobre. Para poder introducirla en la máquina necesitamos una expresión en un lenguaje público, formal y compartido; conceptos intrínsecamente subjetivos, como justicia o felicidad, se distorsionan al traducirse, como explicó Javier Leach en sus reflexiones sobre las diferencias entre los tipos de lenguaje (Leach, 2011).

La ética de las virtudes, situada en el cuadrante *bottom-up* positivo y dependiente de la emergencia de disposiciones morales estables es algo hoy inalcanzable para las máquinas, que carecen de subjetividad (Slote, 1985). Sin embargo, sería posible realizar adaptaciones de este concepto para que las máquinas exhibieran comportamientos análogos a las disposiciones que pudiéramos incluso adaptar dinámicamente según las exigencias del contexto.

Los esquemas basados en reglas afrontan, además, la dificultad de la capacidad de los algoritmos de aprendizaje para eludir restricciones si con ello optimizan su meta (Yampolskiy & Fox, 2013). Para mitigar ese riesgo, en Lumbreras (2017) se introduce la arquitectura *Filtered Decision-Making*: la decisión preliminar del robot se envía a un módulo externo, objetivo e inmutable que verifica su conformidad con el corpus legal; si hay infracción, la acción se bloquea y se activa un modo seguro (Arkin, 2009). Esta pasarela permite actualizar reglas de forma centralizada y evitar sesgos sistémicos, como la reconstrucción de variables sensibles en los modelos de seguros (Avraham et al., 2012). La estructura conocida como *Constitutional ethics*, que se implementó en el chatbot Claude en 2023, es muy similar a esta propuesta: las respuestas de Claude se evalúan con respecto a un conjunto de reglas y, si no las cumplen, la respuesta se retrabaja antes de devolverla al usuario.

5.2.3. La ética del usuario de la IA

Afirma la declaración Antiqua et Nova que «quien utiliza la IA para realizar un trabajo y sigue los resultados crea un contexto en el que él, en última instancia, es responsable del poder que ha delegado» (Dicasterio ..., 2025, nº 46). Es decir, que el uso de IA para la recomendación no exime de la responsabilidad a quien toma la decisión apoyada en la IA.

En el fondo sigue en pie aquella máxima de la ética clásica que afirmaba que había que actuar con conciencia *cierta, recta y tendiendo a que sea verdadera*. Es decir, que por cierta, se debe un asentimiento en el acto sin temor a errar; por recta, se debe ajustar al dictamen de la propia conciencia; y por verdadera, debe estar conforme con el orden objetivo.

Para el caso que nos ocupa del uso de la IA, no se debe delegar todo en la IA sin confirmar, en la medida de lo posible, que no contiene sesgos o que puede causar mal, que está bien alimentada, etc. Es decir que el uso de la IA no exime al humano de garantizar que actúa con conciencia cierta. Esto nos remite a los requisitos de la algorética ya comentados y refuerza que la responsabilidad de las consecuencias de la aplicación de la IA recae sobre el ser humano que la emplea.

Esto incluye la responsabilidad también en la decisión de cuándo y cómo utilizar las herramientas, que debe incluir la prudencia: solo emplearemos la IA cuando nos

sea posible verificar el resultado que nos ofrece contra nuestro conocimiento y experiencia.

Además, la ética del usuario incluye también la protección de nuestra privacidad a cada paso, minimizando los datos compartidos o el respeto a la propiedad intelectual. Por último, se debe destacar la honestidad en la admisión de uso de las herramientas, comunicando cuándo y cómo han formado parte de un proceso creativo o de una decisión.

5.2.4. Ética del trato a la IA

La última de las cuestiones que se plantea a la ética de la IA es la cuestión de si tenemos en la tecnología un *paciente* moral (Coeckelbergh, 2020, pp. 54-59), es decir, qué trato se debe dar a la IA. La cuestión se plantea en línea con el modelo de los círculos de expansión de la consideración moral (Tatay, 2018, p. 401), con la terminología de Peter Singer, que describe una dinámica de extensión progresiva de la preocupación ética que va del interés propio al universo moral completo. El primer círculo abarca el yo individual, donde la motivación predominante es la autopreservación. El segundo incorpora al grupo familiar y de allegados, en el que opera la reciprocidad afectiva. El tercero se proyecta sobre la comunidad próxima–vecindario, organización o nación–y se sustenta en vínculos sociales y normas compartidas. El cuarto engloba a toda la humanidad presente, universalizando la obligación de respetar la dignidad y los derechos de cualquier ser humano, independientemente de la pertenencia grupal. El quinto círculo, finalmente, incluye a todos los seres sintientes, las generaciones futuras y el ecosistema, fundando deberes de justicia intergeneracional y de respeto hacia la vida no humana. Singer sostiene que la razón, al obligarnos a reconocer la arbitrariedad de fronteras anteriores, actúa como motor de esta expansión, transmutando impulsos altruistas locales en principios éticos universalizables. La ampliación de los círculos incluiría la IA, tanto sus algoritmos como los datos en los que se basa.

De manera más clara, plantea esto Luciano Floridi (Floridi, 2013). En su ética *ontocéntrica* se sostiene que el valor moral fundamental no reside exclusivamente en los seres humanos, ni siquiera en los seres sintientes, sino en cuanto algo –sea sujeto, artefacto o proceso– constituye e integra la *infosfera* como entidad informacional; de este modo, todo *objeto de información* (*informational object*) adquiere un estatuto de paciente moral cuya preservación y florecimiento contribuyen al bienestar global del ecosistema informacional. Bajo este paradigma, el criterio normativo deja de ser la maximización de la felicidad humana para convertirse en la promoción de la *infraestructura ontológica*, que posibilita la existencia y la interacción significativa de las entidades informacionales. Ello implica obligaciones de cuidado, conservación y sostenibilidad que se extienden tanto a los entornos digitales (datos, redes, software)

como a los contextos físicos en los que la información se encarna, postulando así un marco deontológico que supera la dicotomía antropocentrismo-biocentrismo, y articula una *ecología de la información* donde el daño o la corrupción de cualquier componente empobrece la totalidad.

Además de la consideración de paciente moral, para algunos, la posibilidad de que en la IA emerja *sintiencia* y consciencia nos debería llevar a diseñar sistemas en los que garanticemos que no pueda generarse sufrimiento en los sistemas y, eventualmente, que pudiéramos reconocerles derechos a los seres que resultasen de este proceso. Por ejemplo, en febrero de 2025, una carta abierta firmada por más de cien expertos, incluido Sir Stephen Fry, alertó sobre los riesgos éticos de desarrollar sistemas de IA potencialmente conscientes. En la carta se argumentaba que estos sistemas podrían experimentar sufrimiento y requerir protección moral. La carta proponía cinco principios rectores: priorizar la investigación sobre conciencia en IA para prevenir el sufrimiento, imponer restricciones al desarrollo de IA consciente, avanzar con cautela y protocolos de seguridad, garantizar la transparencia y el debate público, y evitar afirmaciones engañosas sobre la conciencia en IA.

6. IA Y NUESTRA COMPRENSIÓN DE LA NATURALEZA HUMANA

Los avances de la IA han reforzado la visión dualista del ser humano, del hombre y la mujer, como combinación de una mente que consiste en procesamiento de la infomación y donde reside la identidad, y un cuerpo cuyo único papel relevante es el de proporcionar soporte a estos procesos, reducido a un simple contenedor que puede sustituirse. Esta idea se alimenta tanto de la metáfora del «cerebro-ordenador» como del crecimiento real de las capacidades algorítmicas en la IA (Kurzweil, 2012; Sharp, 2000).

En ese contexto, se haya reforzada también la idea de que la especificidad humana consiste en su capacidad cognitiva de procesamiento de información, que vemos de alguna manera en las interpretaciones tradicionales de la doctrina de la *Imago Dei*, la noción según la cual el ser humano es creado a imagen y semejanza de Dios, y que ha sustentado la dignidad y los derechos de toda persona en las tradiciones judía y cristiana.

Durante siglos la racionalidad fue considerada el núcleo de la *Imago Dei*. Agustín formuló un esquema trinitario memoria-intelecto-voluntad (Agustín, 1991); Tomás de Aquino la situó en el intelecto perfeccionado por la caridad (Tomás, 1962); Lutero y Calvin afirmaron que el pecado la empañó. En el siglo XX, Barth y Brunner desplazaron el acento hacia la relación yo-tú (Brunner, 2014), mientras Ricoeur identificó la imagen con la interioridad reflexiva. Middleton propuso una lectura funcional: el ser humano actúa como virrey de Dios sobre la creación (Middleton, 1994), posición

criticada por la teología de la discapacidad (Eiesland, 1994; Deland, 1999). Aunque la Iglesia Católica, desde el Concilio Vaticano II ha optado por una interpretación existencial o, incluso, otra relacional, alejadas de la comprensión intelectualista de Agustín y Tomás (Amo, 2020).

Las pruebas más conocidas para decidir si una máquina piensa o es consciente –desde el test de Turing hasta desafíos de creatividad o empatía– se formulan como problemas con métricas objetivas. Un algoritmo de aprendizaje por refuerzo puede optimizarlas siempre que disponga de suficientes datos y potencia de cálculo. Pasar la prueba demuestra eficacia funcional, pero no prueba que exista experiencia subjetiva detrás de esa conducta, como explicábamos más arriba.

Para evitar confundir imitación con conciencia, se ha propuesto el *criterio de emergencia*: sólo sería razonable atribuir estados conscientes cuando los signos de subjetividad surgen sin que se los haya fijado como objetivo del entrenamiento. Si el programa aprende a «parecer consciente» porque así lo premia la función de recompensa, seguimos ante un «zombi filosófico», esto es, un sistema que actúa como si tuviera mente, pero sin *qualia* (Lumbreras, 2017).

Las teorías científicas sobre la conciencia refuerzan esta cautela. La Teoría de la Información Integrada (*Integrated information theory*, IIT) sostiene que, para que haya experiencia, debe existir una estructura con una integración causal muy alta, medida por el valor Φ, de modo que el sistema no pueda separarse en partes independientes (Tononi, Boly, Massimini, & Koch, 2016; Koch & Tononi, 2011). Por otro lado, la perspectiva 4E afirma que la mente es encarnada, situada, enactiva y ampliada; depende del acoplamiento sensoriomotor con el entorno y de la práctica corporal en un contexto social (Newen, De Bruin, & Gallagher, 2018). Ambos marcos indican que aún falta por resolver el *symbol-grounding problem*: sin una conexión vivida entre símbolos y mundo, la información que manejan los sistemas actuales carece de significado propio (Harnad, 1990). Mientras les falten esa integración causal y esa corporeidad enactiva, sus logros, por notables que resulten, seguirán siendo puramente computacionales y no fenómenos conscientes.

7. REFERENCIAS BIBLIOGRÁFICAS

Agustín de Hipona. (1991). *Tratado sobre la Santísima Trinidad: Primera versión española, introducción y notas de L. Arias* (Obras de San Agustín, V; B.A.C. 39). Madrid: Biblioteca de Autores Cristianos.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Topdown, bottomup, and hybrid approaches. *Ethics and Information Technology, 7*, 149–155.

Amo Usanos, R. (2007). *El principio vital del ser humano en Ireneo, Orígenes, Agustín, Tomás de Aquino y la antropología teológica española reciente*. Roma: Editrice Università Gregoriana.

Amo Usanos, R. (2020). Una nueva síntesis humanista para un desarrollo humano integral. En J. M.ª Larrú (Ed.), *Desarrollo humano integral y Agenda 2030: Aportaciones del pensamiento social cristiano a los Objetivos de Desarrollo Sostenible* (pp. 143–175). Madrid: BAC. http://hdl.handle.net/11531/45402

Amo Usanos, R. (Ed.). (2023). *Inteligencia artificial y bioética*. Madrid: Universidad Pontificia Comillas.

Arkin, R. C. (2009). *Governing lethal behavior in autonomous robots*. CRC Press.

Asimov, I. (1950). *I, Robot*. Gnome Press.

Avraham, R., Logue, K. D., & Schwarcz, D. (2012). Understanding insurance antidiscrimination laws (Working Paper). SSRN. https://papers.ssrn.com/abstract=2049440

Beorlegui, C. (2023). Implicaciones filosóficas y perspectivas éticas de la inteligencia artificial. En R. Amo Usanos (Ed.), *Inteligencia artificial y bioética* (pp. 19–20). Madrid: Universidad Pontificia Comillas.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Boden, M. A. (2022). *Inteligencia artificial*. Madrid: Turner Publicaciones.

Broncano, F. (Ed.). (1995). *La mente humana*. Madrid: Trotta.

Broncano, F. (1995). Presentación. En F. Broncano (Ed.), *La mente humana* (pp. 1116). Madrid: Trotta.

Brunner, E. (2014). *The Christian doctrine of creation and redemption* (Dogmatics, Vol. 2). Wipf & Stock.

Campbell, R. L., Christopher, J. C., & Bickhard, M. H. (2002). Self and values: An interactivist foundation for moral development. *Theory & Psychology, 12*, 795–823.

Carabantes López, M. (2016). *Inteligencia artificial: Una perspectiva filosófica*. Madrid: Escolar y Mayo Editores.

Corbí, J. E., & Prades, J. L. (1995). El conexionismo y su impacto en la filosofía de la mente. En F. Broncano (Ed.), *La mente humana* (pp. 151-174). Madrid: Trotta.

Cortina, A. (2024). *¿Ética o ideología de la inteligencia artificial? El eclipse de la razón comunicativa en una sociedad tecnologizada*. Barcelona: Planeta.

Cruz, J. (2016). Bioética y teorías de la virtud. En J. J. Ferrer, J. A. Lecaros Urzúa & R. Molins Mota (Coords.), *Bioética: el pluralismo de la fundamentación* (p. 176). Madrid: Comillas.

de Moura, L. M., & Bjørner, N. (2008). Z3: An efficient SMT solver. En C. R. Ramakrishnan & J. Rehof (Eds.), *Tools and Algorithms for the Construction and Analysis of Systems* (pp. 337–340). Springer. https://doi.org/10.1007/9783540788003_24

Dicasterio para la Doctrina de la Fe & Dicasterio para la Cultura y la Educación. (2025). *Antiqua et nova: Nota sobre la relación entre la inteligencia artificial y la inteligencia humana*. Roma: Santa Sede. https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiquaetnova_sp.html

Eiesland, N. L. (1994). *The disabled God: Toward a liberatory theology of disability*. Abingdon Press.

Floridi, L. (2005). Information ethics, its nature and scope. *ACM SIGCAS Computers and Society, 35*(2), 3–8.

Floridi, L. (2013). *The ethics of information*. Oxford University Press.

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

Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial. (2019). *Directrices éticas para una IA fiable*. Bruselas: Comisión Europea. https://digital-strategy.ec.europa.eu/es/library/ethics-guidelines-trustworthy-ai

Handelsman, M. M., Knapp, S., & Gottlieb, M. C. (2009). Positive ethics: Themes and variations. En S. J. Lopez (Ed.), *The Oxford handbook of positive psychology* (pp. 105113). Oxford University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(13), 335346. https://doi.org/10.1016/01672789(90)900876

Herder Editorial S.L. (s. f.). Filosofía de la mente. En *Enciclopedia Herder*. https://encyclopaedia.herdereditorial.com/wiki/Filosof%C3%ADa_de_la_mente

Herder Editorial S.L. (s. f.). Inteligencia artificial. En *Enciclopedia Herder*. https://encyclopaedia.herdereditorial.com/wiki/Inteligencia_artificial

Ibarra, A., & Olivé, L. (Eds.). (2003). *Cuestiones éticas en ciencia y tecnología en el siglo XXI*. Madrid: Biblioteca Nueva.

Jackson, F. (1991). Decisiontheoretic consequentialism and the nearest and dearest objection. *Ethics, 101*, 461–482.

Kant, I. (2002). *Crítica de la razón pura* (M. García Morente, Trad.; J. J. García Norro & R. Rovira, Preparadores). Madrid: Tecnos.

Koch, C., & Tononi, G. (2011, junio). A test for consciousness: How will we know when we've built a sentient computer? *Scientific American, 304*(6), 4447.

Kurzweil, R. (2012). *How to create a mind: The secret of human thought revealed*. Viking.

Leach, J. (2011). *Mathematics and religion: Our languages of sign and symbol*. Templeton Foundation Press.

Lumbreras, S. (2017a). The limits of machine ethics. *Religions, 8*(5), 100. https://doi.org/10.3390/rel8050100

Lumbreras, S. (2017b). Strong artificial intelligence and imago hominis: The risks of a reductionist definition of human nature. En M. Fuller, D. Evers, A. Runehov, & K.W. Sæther (Eds.), *Issues in science and theology: Are we special?* (pp. 157168). Springer.

Lumbreras, S. (2019). *Respuestas al transhumanismo: Cuerpo, autenticidad y sentido* (Vol. 70). Digital Reasons.

Madrid Casado, C. M. (2024). *Filosofía de la inteligencia artificial*. Oviedo: Pentalfa.

Middleton, J. R. (1994). The liberating image? Interpreting the imago Dei in context. *Christian Scholars' Review, 24*(1), 825.

Newen, A., De Bruin, L., & Gallagher, S. (Eds.). (2018). *The Oxford handbook of 4E cognition*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.001.0001

Olivé, L. (2003). Ética aplicada a las ciencias naturales y la tecnología. En A. Ibarra & L. Olivé (Eds.), *Cuestiones éticas en ciencia y tecnología en el siglo XX* (pp. 181–223). Madrid: Biblioteca Nueva.

Olivé, L., & Pérez Tamayo, R. (2001). *Temas de ética y epistemología de la ciencia*. México: Fondo de Cultura Económica.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Padial, J. J. (2019). Técnicas de programación "Deep Learning": ¿Simulacro o realización artificial de la inteligencia? *Naturaleza y Libertad, 12*, 191–210.

Quinlan, J. R. (1990). Learning logical definitions from relations. *Machine Learning, 5*(3), 239–266. https://doi.org/10.1007/BF00117105

Ricoeur, P., & Gingras, G. (1961). "The image of God" and the epic of man. *Cros Currents, 11*(1), 37–50.

Sharp, L. A. (2000). The commodification of the body and its parts. *Annual Review of Anthropology, 29*, 287–328. https://doi.org/10.1146/annurev.anthro.29.1.287

Slote, M. A. (1985). *Commonsense morality and consequentialism*. Routledge & Kegan Paul.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). MIT Press.

Sanguineti, J. J. (2008). Filosofía de la mente. En *Enciclopedia Herder*. https://encyclopaedia.herdereditorial.com/wiki/Filosof%C3%ADa_de_la_mente

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience, 17*(7), 450461. https://doi.org/10.1038/nrn.2016.44

Veruggio, G., Operto, F., & Bekey, G. (2016). Roboethics: Social and ethical implications. En B. Siciliano & O. Khatib (Eds.), *Springer Handbook of Robotics* (pp. 2135–2160). Springer.

Yampolskiy, R., & Fox, J. (2013). Safety engineering for artificial general intelligence. *Topoi, 32*, 217226

En un momento histórico en el que la inteligencia artificial ha dejado de ser una promesa técnica para convertirse en una infraestructura que atraviesa la vida social, este libro ofrece una brújula conceptual y práctica para orientarse en un territorio complejo. La IA se ha convertido en un entorno sociotécnico que reconfigura decisiones, instituciones, relaciones humanas y horizontes culturales. Comprenderla exige, por tanto, algo más que conocimiento técnico: requiere un marco ético capaz de discernir, ordenar y orientar la acción en contextos reales.

Esta obra responde a esa necesidad con una propuesta rigurosa y sistemática. Ofrece una arquitectura conceptual clara —fundamentos, núcleos temáticos y aplicaciones— que permite comprender la ética de la inteligencia artificial como una disciplina práctica, arraigada en los mecanismos reales de la tecnología y en los contextos concretos donde se despliega.

En sus páginas, el lector encontrará, en primer lugar, un sólido "suelo conceptual": los presupuestos tecnológicos, filosóficos y jurídicos que condicionan cualquier evaluación ética. A partir de ahí, el libro identifica los grandes núcleos problemáticos que atraviesan la IA contemporánea, tratados no como categorías abstractas, sino como matrices interpretativas que permiten analizar cualquier sistema. Finalmente, el volumen desciende a los ámbitos de aplicación más mostrando cómo los dilemas éticos adoptan formas específicas cuando se encarnan en prácticas concretas.

El resultado es una obra interdisciplinar que dialoga con la ingeniería, la filosofía, la bioética, el derecho y las ciencias sociales, sin perder en ningún momento la claridad conceptual ni la orientación práctica. Dirigido a investigadores, profesionales y responsables institucionales, pero también a lectores que buscan comprender críticamente el presente, este libro parte de una convicción fundamental: la inteligencia artificial amplía el poder humano, pero no sustituye el juicio moral. Precisamente por ello, cuanto mayor es la capacidad técnica, más urgente se vuelve la exigencia ética.

CÁTEDRA DE BIOÉTICA
COMILLAS CIHS
CÁTEDRA HANA Y FRANCISCO JOSÉ AYALA DE CIENCIA, TECNOLOGÍA Y RELIGIÓN
9 791370 471590