Universitat d'Alacant
Universidad de Alicante
Facultat de Dret
Facultad de Derecho

FACULTAD DE DERECHO

GRADO EN DERECHO

TRABAJO FIN DE GRADO

CURSO ACADÉMICO [2023-2024]

TÍTULO:

¿AGENTES ARTIFICIALES MORALES? CONSIDERACIONES ÉTICAS A PROPÓSITO DE LA INTELIGENCIA ARTIFICIAL

AUTOR:

JULIO DE LUCAS GIL

TUTOR ACADÉMICO:

DR. D. LUCAS MISSERI

RESUMEN

Mediante una revisión bibliográfica cualitativa, se analiza el concepto de inteligencia artificial, pasando por sus distintos enfoques actuales de estudio, así como por los elementos definitorios de los agentes inteligentes. Se prosigue abordando las repercusiones éticas del avance de la técnica en este ámbito —hoy por hoy solo en cuanto a IA débil y específica—, principalmente referidas a la agencia moral y la programación ética de los agentes inteligentes, todo ello en el marco de la ética kantiana y el utilitarismo de Bentham, como modelos al mismo tiempo susceptibles de crítica. A este respecto, siendo la perspectiva crítica de los juristas polacos Brozek y Janik merecedora de particular consideración. Se concluye con la tesis de que no se ha producido aún un desarrollo de la IA tal que permita satisfacer las condiciones requeridas para la agencia moral.

PALABRAS CLAVE: inteligencia artificial, ética, agencia moral, Kant, Bentham

RESUM

Mitjançant una revisió bibliogràfica qualitativa, s'analitza el concepte d'intel·ligència artificial, passant pels diferents enfocaments actuals d'estudi, així com pels elements definitoris dels agents intel·ligents. Es prossegueix abordant les repercussions ètiques de l'avanç de la tècnica en aquest àmbit —ara per ara només quant a IA feble i específica—, principalment referides a l'agència moral i la programació ètica dels agents intel·ligents, tot això en el marc de l'ètica kantiana i l'utilitarisme de Bentham, com a models alhora susceptibles de crítica. En aquest respecte, sent la perspectiva crítica de Brozek i Janik mereixedora de consideració particular. Es conclou amb la tesi que encara no s'ha produït un desenvolupament de la IA que permeta satisfer les condicions requerides per a l'agència moral.

PARAULES CLAU: intel·ligència artificial, ètica, agència moral, Kant, Bentham

ABSTRACT

By means of a qualitative literature review, we analyse the concept of artificial intelligence and its main study approaches nowadays, just as the defining elements of intelligent agents. Then we present the consequences on ethics of technical advances in this area ―only in weak and specific AI for the time being―, mainly related to moral agency and ethical programming, all of that within the framework of Kantian ethics and Bentham's utilitarianism, as models susceptible to being criticised as well. In this respect, deserving Brozek and Janik's critical perspective particular regard. We conclude with the thesis that there is still no development in AI that meets the conditions for moral agency.

KEYWORDS: artificial intelligence, ethics, moral agency, Kant, Bentham

ÍNDICE

I. Introducción

Ya comienza a escucharse de un tiempo a esta parte, con cierta frecuencia, que la humanidad se halla inmersa en la acuñada como «Cuarta Revolución Industrial» —también «Industria 4.0»—. Tras el vapor, la electricidad y la automatización y maquinaria como protagonistas de sus predecesoras, los sistemas ciberfísicos son actualmente el motor de todo un conjunto de innovaciones que han sido calificadas como «tecnologías disruptivas», y que abarcan ámbitos como la conectividad, los datos y el poder computacional, la analítica y la inteligencia, la interacción entre humanos y máquinas, y la ingeniería avanzada (Agrawal et al., 2020).

En medio de estas innovaciones se encuentra la inteligencia artificial —en adelante IA—, como un fenómeno que tras la reciente difusión masiva de tecnologías como GPT-4 (Márquez, 2023) ha sacudido la opinión pública: tanto la calle como las instituciones, políticas, académicas, etc., se han topado con un abanico de consideraciones de muy variada índole —filosóficas, jurídicas, políticas, económicas, p. ej.— que urgen un análisis en profundidad de la cuestión, máxime cuando desde organizaciones supranacionales como la Unión Europea la preocupación ya se refleja en la elaboración de instrumentos normativos —la reciente *AI Act*, aprobada en marzo de 2024 (Parlamento Europeo, 2024)—.

No puede soslayarse —como ya se ha podido afirmar— la relevancia que para el ámbito jurídico tiene la reflexión filosófica acerca de la inteligencia artificial, particularmente desde una perspectiva ética[*]. Algunas cuestiones de esta clase serán analizadas a lo largo del presente trabajo, que tratará de dar una respuesta a las siguientes preguntas:

- ¿Qué es la IA?
- ¿Cómo se comportan actualmente los agentes inteligentes artificiales?
- ¿Cuáles son los elementos definitorios de la agencia moral?
- ¿Podría considerarse a las IA como agentes morales?

[*] En el ámbito iusfilosófico, tal mirada es relevante para los autores positivistas, que entienden la actividad del legislador, de «crear normas», como una centralmente política y/o moral —en contraposición a la de su «aplicación», de índole principalmente técnica y «estrictamente jurídica»—, pero también para otros autores que rechazan tal visión, como los pospositivistas, quienes niegan la existencia de una separación rotunda entre el razonamiento político o moral y el jurídico (Aguiló, 2007).

Para el análisis de tales cuestiones, comenzará abordándose, bajo una óptica generalista, los enfoques de estudio de la IA, el concepto de agente inteligente y las divisiones teóricas de la IA. Seguidamente, previa referencia a los interrogantes éticos que puede suscitar, en torno al concepto de agencia moral se realiza un sucinto repaso a los conceptos de la ética kantiana más pertinentes a este respecto, especialmente el de autonomía, así como al utilitarismo benthamita. Finalmente, se ofrecen diversas observaciones críticas en torno al ajuste a la realidad del comportamiento humano de tales paradigmas filosóficos y un esbozo teórico de la agencia moral, así como su relación con la IA en su actual estado de desarrollo.

II. ¿Qué es la IA?

Lejos de hacer referencia inequívoca a un único concepto, este sintagma encierra una cierta complejidad semántica. Así, a este respecto se han adoptado distintas perspectivas en el estudio de la «inteligencia artificial»

Siguiendo a nuestro entender una ilustrativa explicación (Russell y Norvig, 2022) son cuatro los enfoques, dependiendo de si centran la atención, de un lado, bien en el pensamiento o bien en el comportamiento, y de otro, bien en lo racional o bien en lo humano. Así, puede hablarse de sistemas que:

- Actúan como humanos (*acting humanly*). Se construye una definición operativa de «inteligencia» a partir del test de Turing, propuesto por Alan Turing (1950), para cuya superación es necesario que la computadora reúna capacidades de procesamiento del lenguaje natural, representación del conocimiento, razonamiento automatizado y aprendizaje automático, con la reciente adición a su vez de la visión computacional y la robótica —de acuerdo con el «test de Turing total», una ampliación propuesta por otros investigadores—.

  El «juego de la imitación», planteado bajo la pregunta «¿pueden las máquinas pensar?», consiste en un experimento que involucra a tres participantes: una persona (B), una máquina (A) y un interrogador (C) que permanece situado en una habitación aparte de los otros dos, a quien se encomienda determinar quién es la persona y quién la máquina. Ello tras una conversación de cinco minutos en que C puede realizar diferentes preguntas, ante las cuales tanto A como B —quien ayuda a C a la correcta

identificación de los interrogados— habrán de ofrecer respuestas inteligentes (Turing, 1950).

Según Warwick y Shah (2016), sin embargo, más allá de que la máquina pueda ser razonablemente efectiva en mantener una conversación humana durante un breve período de tiempo, la superación del test de Turing no conlleva de suyo una muestra de inteligencia humana, habida cuenta de la posibilidad que a esta asiste de engañar a su interlocutor —o incluso de mantenerse en silencio ante las preguntas, lo que tampoco permite, al menos teóricamente, una correcta identificación—. Así mismo, ante el desarrollo actual de los «modelos de lenguaje grandes» —o LLM (*large language models*)— como GPT-4, y la creciente capacidad de estos para imitar comportamientos típicos de los humanos, se pone en cuestión la fiabilidad de métodos tradicionales de evaluación de estas tecnologías, no siendo una excepción el mencionado test (Tikhonov y Yamshchikov, 2023).

- Piensan como humanos (*thinking humanly*). La interdisciplinariedad de la ciencia cognitiva permite formular teorías precisas y comprobables sobre la mente humana, aunando los esfuerzos de la simulación informática mediante IA y de técnicas experimentales de la psicología —introspección, experimentos psicológicos, neuroimagen—.

Este rico intercambio de ideas entre disciplinas se ha reflejado en un intenso debate en el seno de las ciencias cognitivas, entre dos posturas —consideradas también paradigmas de la IA—: de un lado, el enfoque tradicional, en cuya virtud la mente es algo similar a una computadora digital procesando un lenguaje simbólico; y, de otro, el enfoque conexionista, basado en un análisis de las capacidades intelectuales a partir de «redes neuronales artificiales». Estas redes constituyen modelos simplificados del cerebro, y se encuentran integradas por un elevado número de nodos, estructurados en una capa de entrada —neuronas «sensoriales»—, una capa de salida —neuronas «motoras»— y una o varias capas ocultas —resto de neuronas—; interconectados a través de un patrón compuesto de ponderaciones y umbrales de activación. (Cameron y Garson, 2019)

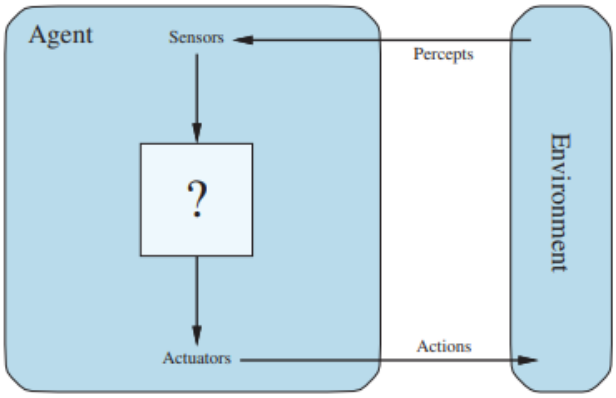
- Piensan racionalmente (*thinking rationally*). A partir de la teoría de la probabilidad y las «leyes del pensamiento» —las reglas de inferencia lógicas—, es posible alcanzar un conocimiento aproximado del mundo y realizar predicciones de futuro, de manera rigurosa y en base a una información incierta; si bien ello no genera de suyo una conducta «inteligente».

- Actúan racionalmente (*acting rationally*). Esperando de los agentes informáticos que operen autónomamente, perciban su entorno, perduren un tiempo prolongado, se adapten al cambio y crean y persigan metas, un «agente racional» es aquel que actúa a fin de lograr el mejor resultado o, ante la incertidumbre, el mejor de los resultados esperados.

i. Los agentes inteligentes

a. La agencia

Siguiendo a estos autores (Russell y Norvig, 2022), puede concebirse el «agente» como algo que percibe —mediante «sensores»— el entorno y actúa —mediante «actuadores»—sobre él, como resulta de la Figura 1:



Nota. Agents interact with environments through sensors and actuators. Tomada de *Artificial Intelligence: A Modern Approach (*p. 55), por S. Russell y P. Norvig, 2022, Pearson Education Limited

Sin embargo, como puede advertirse, esta definición carece de la suficiente precisión para abordar la realidad del objeto. No tan solo puede de ella predicarse su simpleza; también conlleva el riesgo —en tanto en cuanto su amplitud lo permite— de atribuir la condición de agente a numerosos sistemas, incluso cuando en realidad estos no lo son o no deberían haber sido enfocados desde tal óptica (Julián y Botti, 2000).

Se torna preciso, por consiguiente, identificar cuáles son las propiedades de los «agentes» para considerarlos como tales, cuestión tampoco sencilla. Para la elaboración de su teoría, al respecto Wooldrige y Jennings (1995) distinguen entre una noción «débil» y una «fuerte» de la agencia.

Para la primera, acaso la más extendida, se considera agente al *hardware* o —más comúnmente— al *software* que reúne cuatro rasgos:

1. Autonomía, operando sin la intervención directa de los humanos u otros y con un cierto control sobre sus acciones y estado interno (Castelfranchi, 1995, como se cita en Woolridge y Jennings, 1995).

2. Habilidad social, interactuando con otros agentes —y posiblemente humanos— a través de un *agent communication language* (Genesereth y Ketchpel, 1994, como se cita en Woolridge y Jennings, 1995).

3. Reactividad, percibiendo su entorno y respondiendo a tiempo a los cambios en él producidos.

4. Proactividad, mostrando un comportamiento basado en objetivos mediante la toma de iniciativa.

Sin embargo, según afirman Woolridge y Jennings (1995), es de aplicación particular al campo de estudio de la IA un significado más fuerte y específico que el anterior. Para determinados investigadores, un agente es aquel sistema informático que cumple las anteriores condiciones, pero que, siendo estas insuficientes, también ha de caracterizarse en base, entre otras, a nociones ligadas a la mente —conocimiento, creencia, intención y obligación (Shoham, 1993, como se cita en Woolridge y Jennings, 1995)— e incluso a las emociones (Bates et. al, 1992a; Bates, 1994, como se citan en Woolridge y Jennings, 1995).

b. La inteligencia

Tras lo ya expuesto, hacen referencia seguidamente Woolridge y Jennings (1995) a otro atributo —discutido— en el contexto de la agencia: la racionalidad. Este, entendemos, si

se inclina la balanza del lado de los enfoques centrados en lo «racional», no obstante, resulta central.

Volviendo a Russell y Norvig (2022) —que conectan la inteligencia del agente a su racionalidad—, un agente racional debe comportarse seleccionando una acción que, según se espera, maximice su medida de rendimiento dada la información proporcionada por la secuencia de percepciones y su base de conocimientos.

En qué consistan tales *performance measures* —e incluso su pertinencia tal y cómo se conceptúan—, con todo, lejos de constituir una cuestión aséptica, implica la asunción de una teoría ética determinada. Han relacionado los autores, acertadamente, estas con una suerte de «consecuencialismo» (Russell y Norvig, 2022), término inicialmente acuñado por Anscombe (1958), y que, a grandes rasgos, alude a una corriente ética según la cual la bondad o maldad de un acto se determinan ya directamente mediante una comparación entre las consecuencias de ese acto con las de otros actos alternativos, ya mediante un cotejo de «las consecuencias de las reglas, las prácticas o los motivos que determinaron el acto en cuestión con las consecuencias de las reglas, prácticas o motivos que prohíben dicho acto» (Honderich y Trevijano, 2008, p. 220).

Mas no es esta la única caracterización planteada. Poole y Mackworth (2023), por otro lado, consideran que un agente actúa de manera inteligente cuando se adecúa a sus circunstancias y objetivos, así como a sus limitaciones perceptuales y computacionales; toma en consideración las consecuencias tanto a corto como a largo plazo de sus acciones, incluidos los efectos en la sociedad y el entorno; aprende de la experiencia; y es flexible a los cambios en el entorno y en los objetivos.

c. «IA débil» vs. «IA fuerte»

En un campo tan amplio como este, son múltiples las tipologías y clasificaciones establecidas. Con la publicación del artículo *Minds, brains and programs* en el año 1980, sin embargo, introduce el filósofo John Searle —especializado en la mente y el lenguaje— una distinción fundamental para la IA, sobradamente extendida en la actualidad, entre un sentido débil y otro fuerte de la misma.

Para el autor, mientras que en la primera —*weak AI*— el principal valor de la computadora se basa en proporcionarnos una herramienta muy poderosa para el estudio de la mente, en la segunda —*strong AI*—, más allá de esto, la computadora

adecuadamente programada es una «mente», en la medida en que se afirma que, dados los programas correctos, puede «entender» y tener otros estados mentales.

A partir del conocido experimento mental de «la habitación china» Searle (1980) dirige una crítica contra los propósitos de la investigación científica en IA, concretamente respecto de aquella que califica de «fuerte»; y con ello también contra la tesis en cuya virtud los procesos mentales son procesos computacionales sobre elementos formalmente definidos, a la que hace frente.

Este problema (López de Mántaras, 2018) entronca a su vez con la distinción entre la «IA específica», cuando esta es concebida para tareas concretas —por ejemplo, jugar al ajedrez al nivel de Gran Maestro— y la «IA general», de carácter complejo y asimilable, por ello, a la propiamente humana. Si bien alcanzar esta segunda es el objetivo último de la IA, en la medida en que la inteligencia humana surge a partir de interacciones particulares proporcionadas por el cuerpo —especialmente los sistemas perceptivo y motor—, así como de los procesos de socialización y culturización—, difícilmente podría lograrse esta misión en el caso de la IA.

Siguiendo a López de Mántaras (2018) cabe resaltar, en otro orden de cosas, que esta empresa —ambiciosa cuando menos— no resulta incompatible con la inviabilidad de la IA fuerte. Así, se afirma que «toda IA fuerte será necesariamente general pero puede haber IA generales, es decir, multitarea, que no sean fuertes, que emulen la capacidad de exhibir inteligencia general similar a la humana, pero sin experimentar estados mentales» (p. 46).

Partiendo de la hipótesis de la «singularidad» —la radical expansión de la inteligencia humana a partir de su fusión con formas no biológicas—, bajo la idea de un desarrollo tecnológico de carácter exponencial, a la luz de un examen histórico —frente a una visión lineal, más intuitiva—, defendida por Kurzweil (2005), también es corriente la mención a la «superinteligencia».

Esta es definida por Bostrom (2016) como «cualquier intelecto que exceda en gran medida el desempeño cognitivo de los humanos en prácticamente todas las áreas de interés» (p. 22); quien además distingue diversas formas de superinteligencia —de velocidad, colectiva y de calidad (en cierto sentido práctico, equivalentes)—, aunque reconociendo, sin embargo, que aún la cognición artificial se sitúa en un escalón inferior

con respecto a la humana; a falta ello de un «despegue», una «explosión de inteligencia» dada a partir de la comprensión de su funcionamiento, por parte de una IA «seminal», que le permita adquirir la capacidad de producir nuevos algoritmos y estructuras computacionales, en un proceso de automejoramiento sostenido.

A la fecha, empero —y como ya se desprende de lo expuesto—, el avance de la técnica únicamente permite hablar de IA débiles y específicas.

ii. La dimensión ética de la IA

Hasta el momento, se han abordado los distintos enfoques teóricos de la IA, la naturaleza de los agentes inteligentes y los avances en su desarrollo, en varias etapas. Pero cabe preguntarse: ¿cuáles son las implicaciones éticas de tales planteamientos? Hemos de tratar con urgencia una nada fútil cuestión: si puede alumbrar la IA «agentes morales».

Según una concepción «estándar» de tal noción —normativa—, puede convenirse que un agente es moral si y solo si: (1) Tiene obligaciones o requerimientos morales, y cuenta con las capacidades para (2) actuar libremente —con autonomía—, (3) deliberar sobre los requerimientos y (4) entender y aplicar correctamente reglas morales en casos paradigmáticos (Himma, 2009; Monasterio, 2019).

Ello permite diferenciar entre los «agentes» morales, a quienes se imponen determinados deberes y obligaciones; y los «pacientes» morales, como acreedores de al menos un deber u obligación. Sin embargo, los agentes morales son a menudo —cuando no siempre— pacientes morales: todos los humanos adultos son pacientes morales. Ahora bien, existen también, en términos morales, muchos pacientes que no son agentes —así sucede respecto de los niños recién nacidos, si bien previsiblemente adquirirán en el futuro también la condición de agencia—. Tal distinción no es espuria, en la medida en que la corriente estándar liga conceptualmente la agencia moral con la responsabilidad por el propio comportamiento (Himma, 2009).

Un concepto clave aparece en la discusión: la «autonomía»; término que ya ha sido anteriormente mencionado, como propiedad de los agentes para Woolridge y Jennings (1995). Apuntando a una conceptualización más en detalle, Castelfranchi y Falcone (2003) definen esta, en sus rasgos más básicos y en términos relacionales, a partir de tres entidades: el sujeto principal (X), cuya autonomía es analizada y/o evaluada; una función/acción/objetivo (μ), que debe ser realizada o mantenida por tal sujeto principal;

y un sujeto secundario (*Z*) —o una pluralidad de ellos— con respecto al cual el sujeto principal debería ser considerado autónomo dada la función/acción/objetivo objeto de especificación. Entre otras clases de autonomía expuestas, se ocupan también estos autores de la «autonomía de objetivos» —*goal autonomy*—: en este sentido, sería *X* un sujeto autónomo tanto cuando tiene su propio objetivo a alcanzar como cuando adopta una tarea τ dada por el agente *Z*; esto es, bien rechace o bien acepte las peticiones de otro, lo determinante es que se guíe por sus propias razones.

Ahora bien, por más que pueda producirnos cierta sensación de familiaridad ¿es esta noción —fruto de una necesaria «operacionalización» para a su dominio, según Castelfranchi y Falcone (2003)— útil a los propósitos de una reflexión en materia ética? Qué pueda ser o no considerado un agente moral y cómo estos, para ello, hayan de actuar, es un asunto que ha de plantearse en términos distintos.

Ligado a la agencia moral, en el contexto de una «ética de las máquinas», es preciso acometer a su vez el análisis de las principales aproximaciones en cuanto a la programación ética de los agentes artificiales, a saber: de arriba hacia abajo (*top-down*), por una parte; y de abajo hacia arriba (*bottom-up*), por otra; si bien en la práctica los ingenieros emplean ambos enfoques. Mientras que el primero implica tomar como referencia una teoría ética general previamente especificada, así como el análisis de los requisitos computacionales para guiar el diseño de algoritmos y subsistemas aptos para su implementación, el segundo trata los valores normativos como algo implícito en la actividad de los agentes, en vez de explícitamente articulado —o articulable— en torno a una teoría general (Wallach et al., 2008).

Entre aquellas «teorías éticas generales» se hallan concepciones procedentes de fuentes muy variadas —religión, filosofía, literatura, etc.—, que ofrecen ejemplos como la «regla de oro», los Diez Mandamientos, las éticas consecuencialistas y/o utilitaristas, el imperativo categórico kantiano y otras éticas del deber, el ordenamiento jurídico, las virtudes aristotélicas o las «tres leyes de la robótica» de Asimov (Wallach et al., 2008).

Tanto el deontologismo como el consecuencialismo son susceptibles de formalizarse lógicamente, mediante una heurística de implementación (Monasterio, 2009), por lo que los analizamos seguidamente, tomando como concepciones ilustrativas, respectivamente, el pensamiento filosófico de Immanuel Kant (1724-1804) y de Jeremy Bentham (1748-1832).

III. La ética kantiana

Tras la publicación en el año 1781 de la *Crítica de la Razón pura* —obra con que, fruto de «una gran iluminación» recibida en 1769, el filósofo prusiano lleva a cabo la «revolución copernicana» en su pensamiento, que le permite superar el racionalismo y el empirismo, el dogmatismo y el escepticismo (Reale y Antiseri, 1988)— verá la luz la *Fundamentación de la Metafísica de las Costumbres* —FMC— (1785) y posteriormente la *Crítica de la Razón Práctica* —CRP— (1788) obras donde se desplegará la teoría ética del autor.

A. El fin de la razón (práctica)

«Ni en el mundo, ni, en general, tampoco fuera del mundo, es posible pensar nada que pueda considerarse como bueno sin restricción, a no ser tan sólo una buena voluntad.» (FMC, 1977, p. 27).

A partir de estas líneas, que inauguran el capítulo primero de FMC, Kant sienta una premisa básica a partir de la cual construirá toda la reflexión que sigue: el valor absoluto de la mera voluntad, como bien supremo y condicionante de la obtención de otros —v. gr. la felicidad—. Y no será el fin de la facultad práctica de la razón otro sino la producción de una voluntad «buena en sí misma» (FMC, p. 32).

Para determinar en qué consiste esta, recurre el filósofo a dos conceptos nucleares: el imperativo y el deber; que a continuación tratamos.

De un lado, por «imperativo» entiende aquel principio objetivo —en contraposición a la «máxima», de carácter subjetivo— en cuya virtud «manda» la razón una acción u omisión determinada (FMC, pp. 39, 72). Incluidos en tal género se hallarán tanto los imperativos «hipotéticos», que la ordenan ya para un fin posible —problemáticos, «reglas de la habilidad»—, ya para un fin real —asertóricos, «consejos de la sagacidad»—, como el imperativo categórico —apodíctico—, aquel que no condiciona la realización de la conducta a la obtención de propósito alguno (FMC, p. 62) y que, por consiguiente, alcanza el verdadero rango de ley moral (FMC, pp. 64-65).

De otro, identifica el «deber» con la «necesidad de una acción por respeto a la ley» (FMC, p. 38). Pero no se refiere Kant a un mero obrar «conforme al deber». Sucede en ocasiones que el sujeto, a la hora de actuar, se ve sometido a las exigencias de ciertas inclinaciones

que operan como «resortes» en el terreno empírico (FMC, pp. 33-34). De los fines subjetivos, sin embargo, no es posible extraer principios prácticos elevables a ley (FMC, p. 82). El imperativo categórico ordena actuar de manera inmediata e incondicionada, no atendiendo al objeto de la acción, a sus efectos *a posteriori*, sino a la acción en sí misma considerada.

A diferencia de lo que sucede con los imperativos técnicos y pragmáticos ya referidos, el imperativo moral supone, por consiguiente, una proposición sintética-práctica *a priori* (FMC, p. 123). Únicamente cabe atribuir, por tanto, valor moral a aquellas acciones realizadas «por deber» y sin afectación por parte de condiciones subjetivas, ancladas en la experiencia (FMC, pp. 37-39).

Tan solo repara Kant en un sentimiento, al que tilda de «moral» (FMC, p. 132), si bien le concede un particular estatus en su esquema: el «respeto», que entiende como un efecto de la ley sobre el sujeto, cuando este es consciente de la determinación inmediata de la voluntad por parte del imperativo categórico (FMC, p. 40). Queda a salvo, luego, lo afirmado anteriormente: una buena voluntad no se determina más que por las condiciones objetivas proporcionadas por la razón —a cuya obtención esta se ordena— (FMC, p. 59).

La «felicidad», referida a un máximo de bienestar en el estado actual y futuro, se trata, no obstante, de un concepto indeterminado, un ideal «de la imaginación» —no «de la razón»— (FMC, pp. 67-68), el cual, como ya se ha dicho, no sirve de fundamento a la moralidad, sino más bien lo contrario: contribuye a situar en una misma clase los motivos productores de virtud y de vicio de tal modo que, por medio de una operación de «cálculo» —repárese en este término—, queda eliminada la diferencia entre uno y otro (FMC, p. 104).

En esta línea, contrariamente a lo defendido desde otras concepciones éticas —como el utilitarismo benthamita, que tendremos ocasión de analizar más adelante—, no se nos presenta la moral como una «doctrina de la felicidad» —una enseñanza dirigida a alcanzar esta—, sino como una guía para convertirse uno en «digno» de la misma: respecto de ella, la moral —siguiendo una idea ya apuntada—opera como una condición racional *sine qua non*, más no como un medio para adquirir la felicidad (CRP, p. 161).

B. La noción de «autonomía»

Una vez ha enunciado el filósofo las tres conocidas formulaciones del imperativo categórico, procede a concretar cuál es la relación entre ellas, que se reconduce a la autonomía de la voluntad, a la que se refiere como «principio supremo» (FMC, p. 101) y «universal» de la moralidad (FMC, p. 121).

A este concepto, que queda condensado en el fórmula «obra según una máxima que contenga en sí al mismo tiempo su validez universal para todo ser racional» (FMC, p. 97), llega Kant a través de la libertad. Siguiendo el razonamiento previamente expuesto, señala la imposibilidad de concebir una razón que no se tenga a sí misma como libre, autora de sus principios, con independencia de los influjos sensibles ajenos a ella. (FMC, p. 114). En la medida en que el ser racional es consciente de su propia naturaleza, perteneciente tanto al mundo sensible como al mundo inteligible, surge la conciencia del deber de actuar conforme al imperativo categórico (FMC, pp. 122-123).

Así, aparece la «libertad», entendida, de un lado, en un sentido negativo, como la ausencia de determinación de la voluntad por parte de causas sensibles, y de otro, en un sentido positivo, como facultad de obrar conforme al dictado de la causa racional derivada de la condición de validez universal de la máxima como ley (FMC, p. 129). En palabras del filósofo, «voluntad libre y voluntad sometida a leyes morales son una y la misma cosa» (FMC, p. 112).

Se erige en el terreno ético, por consiguiente, la «autonomía» frente a la «heteronomía de la voluntad», que brota —esta última— cuando la voluntad, tratando de determinar la ley, «sale de sí misma a buscar esa ley en algunos de sus objetos» (FMC, p. 102), en la «sensibilidad» (FMC, p. 133). Carece de una posición antagónica con respecto a la autonomía moral, en cambio, la heteronomía de las leyes de la naturaleza, de que ya hemos hecho somera mención, derivada de la inserción en el mundo sensible.

Ahondando en la dicotomía, posteriormente en la CRP, Kant relacionará algunas teorías éticas basadas en la heteronomía de la voluntad, distinguiendo, a propósito de los motivos materiales determinantes del obrar entre: «subjetivos», extrínsecos —de la educación (Montaigne), de la constitución civil (Mandeville)— e intrínsecos —del sentimiento físico (Epicuro), del sentimiento moral (Hutcheson)—; y «objetivos», a su vez también

extrínsecos —de la voluntad de Dios (Crusius y otros moralistas teológicos)— e intrínsecos —de la perfección (Wolff)— (CRP, pp. 59-62).

Pese a que podría tildarse de fantasiosa la teoría ética kantiana —dada la ausencia de ejemplos, palpables en la práctica, de comportamientos «por deber», cuando no quizás directamente su inexistencia—, es consciente de esto el filósofo, quien nos recuerda dos cautelas en que hemos de reparar (FMC, pp. 50-53), las cuales creemos conveniente sistematizar como sigue: de un lado, el carácter normativo de la teoría, en cuya virtud no persigue esta una descripción de cómo «se comportan» los sujetos en la práctica, sino cómo «deben comportarse»; y, de otro, su concepción del sujeto ético, no reducida al ser humano, sino comprensiva del conjunto de los «seres racionales en general». Llega a predicar incluso de la «autonomía» su consideración como «fundamento de la dignidad de la naturaleza humana y de toda naturaleza racional» (FMC, p. 94).

IV. El utilitarismo de Bentham

Pese a no tratarse, en rigor, del iniciador del utilitarismo, lo cierto es que, a nivel sistemático, es Bentham quien funda este como teoría ética bajo una preocupación de reformar la sociedad, lo que lo lleva a publicar, en el año 1789, su conocida obra *An Introduction to the Principles of Morals and Legislation* —en adelante IPML— (Lazari-Radek y Singer, 2017), donde sienta las bases de un utilitarismo que bien puede considerarse un «hedonismo psicológico», mas con un eminente giro práctico —el empirismo inglés clásico se había ocupado de la naturaleza, extensión y límites del conocimiento humano— (Copleston, 1993).

Si bien probablemente no se trata de una toma de partido consciente entre dos corrientes, —a juzgar por el contexto en que se inserta—, con frecuencia se ha calificado el benthamismo de un «utilitarismo del acto», frente a otro utilitarismo, que podríamos denominar «de la regla». Para el primero, un acto es correcto si y solo si resulta al menos en tanto bienestar general como cualquier acto que podría haber llevado a cabo el agente; mientras que, para el segundo, un acto es correcto si y solo si viniera permitido por un sistema de reglas cuya aceptación general resultara en tanto bienestar general como la aceptación general de cualquier otro sistema de reglas (Eggleston, 2014).

A. El principio de utilidad

La naturaleza ha colocado a la humanidad bajo el gobierno de dos maestros soberanos: el dolor y el placer. A ellos solos les corresponde señalar qué debemos hacer, así como qué haremos. De un lado, el estándar de corrección e incorrección, de otro, la cadena de causas y efectos, están fijados a su trono (IPML, p. 11).

Así las cosas, Bentham ofrece una perspectiva descriptiva acerca de la moral humana, basada en la idea de que todo ser humano por naturaleza busca el «placer» y evita el «dolor» —estos conceptos tal y como aparecen en la opinión común—, mas, además de este hedonismo, no rechaza la pretensión de establecer un criterio objetivo de moralidad de las acciones humanas, en base al principio de utilidad (Copleston, 1993).

El «principio de utilidad» —o «de mayor felicidad», una denominación preferible para Bentham— permite la aprobación —o reprobación— de cualquier acción, según si de ella se deriva una tendencia al aumento —o disminución— de la felicidad de la parte cuyo interés está en discusión, sea esta bien un individuo particular o bien la comunidad en general. Todo ello entendiendo la «utilidad» como la propiedad de cualquier objeto por cuyo intermedio este tiende a producir beneficio, ventaja, placer, bien o felicidad, y a prevenir daño, dolor, mal o infelicidad (IPML, p. 11-12).

Por lo que hace a hace a palabras como «deber», «correcto» o «incorrecto», y a su interpretación, al calificar una acción «conforme con el principio de utilidad», de esta se predica que debe realizarse o que, como mínimo, no se identifica con una que no deba realizarse; en términos similares, una acción llevada a cabo de acuerdo con tal principio es correcta o, por lo menos, no es incorrecta (IPML, p. 13).

Reconoce Bentham, no obstante, cómo es este principio en raras ocasiones perseguido de manera consistente por las personas (IPML, p. 13). Esta inaplicación, empero, no conlleva su incorrección, que él mismo se ocupa de combatir a través de distintos argumentos (IPML, p. 15-16), así como tampoco la corrección de otros principios considerados «adversos», como pueden ser ora el «principio de asceticismo», que determina la aprobación —o reprobación— de una acción en la medida en que tienda a la disminución —o aumento— de la felicidad (IPML, pp. 17-18), ora el «principio de simpatía», en cuya virtud una acción es aprobada —o reprobada— en función de los sentimientos internos

de aprobación —o reprobación— en sí mismos considerados, sin necesidad de atender a razón extrínseca alguna (IPML, pp. 21-25).

B. El cálculo «felicífico»

Tratándose los «placeres» y los «dolores» de instrumentos para la acción, es necesario entender cuál es su fuerza en cada caso, esto es, el valor de estos (IPML, p. 38).

Por lo que a ellos respecta, para cada persona «por sí misma» Bentham hace depender el valor de los placeres y dolores «por sí mismos» de cuatro circunstancias: su intensidad; su duración; su certeza o incertidumbre; y su proximidad o lejanía (IPML, p. 38). A su vez, cuando se trata de estimar la tendencia de un «acto» cabe tomar en consideración adicionalmente: su fecundidad, como posibilidad de ser sucedido por sensaciones del mismo signo —placeres o dolores—; y su pureza, como posibilidad de no anteceder a sensaciones del signo opuesto —bien dolores tras placer, o bien placeres tras dolor—. Y, por último, siendo el sujeto cuyo interés se halla en discusión una pluralidad de personas, una séptima y última circunstancia: el alcance del acto, es decir, el número de afectados por la extensión del mismo (IPML, p. 39).

Aunque no se espere, hacia este proceso, un recta adhesión, que suponga su estricta aplicación con carácter previo a todo juicio moral u operación judicial o legislativa (IPML, p. 40), con todo, la tendencia general de un acto —más o menos perniniciosa— se hará de depender de la suma total de sus consecuencias, esto es, de la diferencia entre las buenas y las malas, mas únicamente tomando en cuenta las consecuencias «materiales», aquellas que bien consisten en un dolor o un placer o bien influyen en su producción (IPML, p. 74).

Pese a afirmar que el benthamita se trata de un pensamiento hedonista, ello no es óbice para que, dentro de los placeres y los dolores —que el autor caracteriza como «percepciones interesantes» y divide en simples y complejas (según si pueden volverse adicionalmente en más o no, respectivamente) (IPML, p. 42)— se refiera, como placeres y dolores simples, entre otros, a los de «benevolencia» —también de «buena intención» o de los «afectos sociales»—, resultantes , en uno y otro caso, de observar cualquier placer siendo poseído por los seres objeto de benevolencia, es decir, los seres sensitivos, y a su vez de observar cualquier mal siendo soportado por aquellos mismos (IPML, pp. 44-48);

lo que permite excluir la consideración del ser humano como alguien «por naturaleza necesariamente egoísta o autosuficiente» (Copleston, 1993).

C. La valoración de las acciones

Si bien lo hace originalmente el filósofo con una perspectiva punitiva, ofrece siete elementos de análisis, en términos éticos, de una conducta, a saber (IPML, pp. 74-75): el acto en sí mismo —aquello que se realiza—; las consecuencias en que se realiza; la intencionalidad que lo puede seguir; la conciencia, inconsciencia o falsa conciencia que lo puede acompañar; los motivos de los que ha surgido; y la disposición general que lo indica; de los cuales tres serán objeto de una sucinta explicación seguidamente.

En primer lugar, puede ligarse la «intencionalidad» al acto en sí mismo y/o a sus consecuencias, siendo la acción enteramente intencional cuando atiende a ambos elementos e «inintencionada» cuando no guarda relación con ninguno de ellos (IPML, p. 84). La intención dependerá de cuál sea el estado de la voluntad o intención respecto del acto en sí mismo y el estado del entendimiento —o de las facultades perceptivas— de las circunstancias que lo rodean, existiendo, en relación con ellas, tres supuestos: la «conciencia», cuando se conocen de modo preciso tales circunstancias; la «falsa conciencia», cuando erróneamente se concibe su existencia; o la «inconsciencia», cuando se yerra en percibirla (IPML, p. 75).

En segundo lugar, y en un lugar destacado, hemos de referirnos a los «motivos» como todo aquello que influye en la voluntad de un ser sensible, que lo conduce a actuar o a abstenerse de hacerlo: los motivos «prácticos», del obrar —distintos de los «especulativos», carentes de influencia externa por circunscribirse a la esfera intelectual— (IPML, pp. 96-97). No existen, sin embargo, motivos intrínsecamente malos o buenos: que lo sean, en todo caso, dependerá de los efectos que produzcan, de su tendencia a producir o a evitar placeres o dolores (IPML, p. 100).

En sentido estricto, por consiguiente, de nada puede decirse que sea bueno o malo, salvo en sí mismo —válido únicamente para los dolores y/o placeres— o en atención a los efectos que produce —en términos de dolor y/o placer—. Quizás en un sentido más laxo, también puede estimarse la bondad o maldad de un acto a partir de la intención que subyace al mismo, con referencia esta bien a los motivos que la causan o bien a las consecuencias del acto que tiene por objeto, siendo estas últimas —las consecuencias—

finalmente buenas o malas, no obstante, a partir de las circunstancias en que se produzca el acto —únicamente objetos del entendimiento, y no de la voluntad— (IPML, pp. 88-89).

Es posible aseverar, por ende, que ninguna acción para Bentham puede propiamente reputarse como desinteresada, dada la «causación» de todas ellas por la anticipación, por parte del individuo, de los placeres y los dolores que constituyen para él la percepción de su interés (Crimmins, 2014).

V. La crítica de Brozek y Janik

Con la publicación del artículo *Can artificial intelligences be moral agents?*, Brozek y Janik (2019), se proponen el análisis de dos importantes teorías ético-filosóficas, el kantismo y el utilitarismo —por su preeminencia en las ciencias morales contemporáneas y suponer un ejemplo de la inatención de la dimensión psicológica en los debates morales—, la delimitación de la estructura de la moral humana, y, por último y cómo propósito principal, la hipotética consideración de las máquinas como agentes morales.

Por cuanto respecta al *homo kantianus*, señalan, se trata de una «criatura altamente idealizada», y ello por dos motivos: en primer lugar, la elevada complejidad de un hipotético sistema de reglas morales apto para regir todas las posibles circunstancias con deberes *prima facie* en conflicto, «una tarea cognitiva imposible» para el ser humano; y, en segundo lugar, la profunda discordancia entre los resultados de la investigación psicológica sobre la acción moral humana con el obrar «por deber», al margen de cualquier emoción, mandado por el imperativo categórico kantiano (Brozek y Janik, 2019).

Esta distancia se estrecha, empero, al intercambiar en la relación, el *homo sapiens*, como término de comparación, por un agente artificial: sí resulta posible, para el sistema de inteligencia artificial, actuar únicamente por deber, en contraposición al ser humano. En términos cognitivos, respecto al manejo de aquel complicado trasunto de reglas morales, es patente la superioridad de los agentes artificiales con respecto a las habilidades humanas. Además, pueden teóricamente ignorar en su obrar las emociones, como debe suceder para el *homo kantianus* —si bien ello también excluye el «respeto», sentimiento

particular al cual ya nos hemos brevemente referido, propio incluso de los seres racionales que actúan moralmente— (Brozek y Janik, 2019).

En cuanto al ajuste del modelo teórico a la realidad, otro tanto ha de indicarse respecto del *homo benthamus*, capaz de llevar a cabo un cálculo de la utilidad total resultante de una acción concreta a partir de un análisis de todas sus consecuencias —algunas de ellas alejadas e inciertas—, y de llegar a mantener al margen el propio interés y/o determinadas emociones cuando la elección que proporciona la mayor utilidad así lo exige; si bien igualmente en este caso el desempeño de los agentes artificiales, en el aspecto cognitivo y motivacional, resulta mayor (Brozek y Janik, 2019).

Dada la carencia, por parte del *homo sapiens,* de habilidades cognitivas y mecanismos motivaciones asimilables, aún en lo aproximado, a los propios tanto del *homo* kantianus como del *homo benthamus,* frente a tales concepciones, Brozek y Janik (2019) adoptan, en su exposición sobre la estructura de la moral, una postura descriptiva, no normativa, de qué significa actuar moralmente; y ello ofreciendo tres ideas principales:

- «Tesis de la intuición»: las decisiones moralmente relevantes, en su mayoría, se adoptan de modo inconsciente, a través de una intuición impulsada por las emociones, entrenada mediante las interacciones con otras personas, pero no exclusivamente moral—.

- «Tesis de la comunidad»: no existen comunidades sin moral, en la medida en que esta beneficia al grupo y a sus miembros facilitando la cooperación y la convivencia, las prácticas morales integran un todo social más amplio, y los agentes morales deben ser reconocidos como tales, en el marco de una cultura dada.

- «Tesis del andamiaje»: las teorías morales se alcanzan mediante una abstracción de las prácticas morales realmente existentes en las interacciones sociales e interactúan con tales prácticas a través de un circuito de retroalimentación, moldeando la intuición moral a largo plazo, pese a que no puedan identificarse con las mismas.

A juicio de Brozek y Janik (2019), no se trata la agencia moral de un concepto absoluto, caracterizado por un conjunto estricto de criterios de determinación. Son dos los tipos de condiciones: una externa y otra interna.

Por una parte, en su vertiente «externa», el pretendido agente moral —un individuo o incluso un objeto inanimado— requiere de «reconocimiento» en el seno de una comunidad, con la correlativa antropomorfización que supone la adscripción a él de determinadas habilidades cognitivas y motivacionales con la acción humana como referencia.

Por otra, en su vertiente «interna», aquella se estructura en capas interconectadas y sucesivas, que permiten la distinción entre agentes morales:

- «Irreflexivos» —o «superficiales»—, que obedecen acríticamente, por medio de la intuición, las reglas morales de la comunidad, para hacer encajar su conducta en los parámetros de corrección imperantes —los niños pequeños—.

- «Reflexivos», con capacidad para la justificación de su conducta y, en algunas ocasiones, también para el razonamiento moral en términos abstractos —la mayoría de los adultos—.

- «Sofisticados» —o «profundos»—, con un nivel de autoconciencia moral que les permite actuar con un mayor grado de abstracción en la determinación de su obrar, ponderando distintas razones —únicamente ciertos individuos—.

Estos dos tipos de condiciones, para Brozek y Janik (2019), externas e internas, que interactúan entre sí, son individualmente necesarias, si bien así consideradas insuficientes a los efectos de otorgar aquella condición. Luego puede existir un reconocimiento antropomorfizante sin apoyo en determinadas habilidades cognitivas y motivacionales, pero también una ausencia del mismo aun a pesar de su concurrencia; tanto en uno como en otro caso, no dándose ambas conjuntamente, no podrá hablarse de un verdadero «agente moral». Cabe preguntarse cuál es el encaje de los agentes artificiales en este esquema: ¿pueden convertirse en agentes morales?

La respuesta que dan al problema es nítida: tal y como actualmente se halla conceptualizada la agencia moral, la vigente arquitectura de la IA —por muy compleja y

sofisticada que pueda llegar a ser en ciertos supuestos —no puede satisfacer mínimamente las condiciones referidas. Externamente, el reconocimiento de la agencia tiene como referencia al humano y su inserción en prácticas morales complejas, dependientes de la cultura y basadas en una intuición comunitaria. Internamente, más allá de la incorporación de «módulos emocionales», no actúan las emociones como motores del comportamiento de los agentes artificiales. Sentencian la cuestión afirmando que cuanto más se aproximen las máquinas a los humanos, esto es, sean capaces de participar plenamente en prácticas morales, más probable será a su vez el reconocimiento como agentes morales (Brozek y Janik, 2019).

VI. Conclusión

Partiendo de las distintas consideraciones apuntadas, se advierte con toda claridad un hecho: la IA no constituye una realidad única; más bien se trata de un fenómeno complejo, objeto de estudio desde distintas disciplinas y bajo enfoques diversos, en cualquier caso expresivos de este rasgo y, en muchos casos, complementarios entre sí.

Si bien tan solo puede hablarse, en la actualidad, de IA débiles y específicas, un potencial desarrollo tecnológico en este campo durante los próximos años —con diferentes perspectivas, más o menos ambiciosas (que alcanzan incluso a la hipótesis de la «superinteligencia»)— aconseja examinar cuáles son las implicaciones tanto en su programación ética como a propósito de la agencia moral —anidada esta en ocasiones a la responsabilidad por la propia conducta— de los «agentes inteligentes».

Se ofrecen distintas acepciones «técnicas» de tal agencia: una básica, acaso demasiado amplia en cuanto a su referente; acompañada de otras dos, una débil, que ya incorpora un elemento de «autonomía» —más asemejado, empero, a la libertad en un mero sentido negativo—, y una fuerte —quizás algo más interesante, por cuanto se refiere a cuestiones ligadas a la mente o las emociones—. Asimismo respecto del término «inteligencia», relacionado en mayor o menor medida con una racionalidad nada aséptica en términos ético-filosóficos por su ligazón a unos determinados indicares de desempeño.

Todo ello para resultar en un complejo, el «agente inteligente», que con gran predicamento puede de este modo nominarse en el terreno de la ingeniería, de las ciencias

de la computación, pero cuya agencia, en términos morales, no supone una cuestión sencilla ni mucho menos indubitada.

Analizar la noción estándar —normativa— de la agencia moral nos conduce por fuerza a la discusión acerca de la programación ética de los agentes inteligentes. Si para atribuir a un sujeto la cualidad de agente moral ha de aunar este la capacidad para actuar con autonomía y razonar sobre sus requerimientos morales, harto complicado se nos antoja este propósito sin contar con aquella programación.

Tal programación puede diseñarse mediante un enfoque «de abajo hacia arriba» o «de arriba hacia abajo», como sucede —en este último caso— cuando se plantean como posibles modelos de comportamiento para las IA el formalismo kantiano o el utilitarismo benthamita, reconociblemente antitéticos, como se refleja en la llamativa posición que ocupa la felicidad en uno y otro sistema: en el primero, apriorístico, marginada en pos de la búsqueda de la buena voluntad, como fin de la razón práctica y fundamento de la moralidad, y regida ella conforme al imperativo categórico —al margen, por tanto, de influjos sensibles, como las emociones—; mientras que elevada a principio supremo en el segundo, donde la bondad o maldad de un acto pasa a determinarse en base a un cálculo de dolores y placeres —profundamente anclado en lo sensible—.

Bajo el ánimo de argumentar en favor de una concepción de la agencia moral desde el comportamiento humano, resultan estas teorías éticas, no obstante, poco realistas en cuanto a la verdadera estructura de la moralidad humana —a pesar de un posible mayor ajuste en el caso de la IA—, lo que lleva a la necesidad de reconfigurar aquella desde una perspectiva tanto interna —presencia de habilidades cognitivas y mecanismos motivacionales— como externa —reconocimiento en el seno de una comunidad—, como realizan Brozek y Janik.

En la actualidad, como ya se ha tenido ocasión de afirmar, la IA, en sus diferentes manifestaciones, se halla en un estadio de desarrollo en que, desde tal entendimiento, los «agentes inteligentes» no pueden ser a su vez considerados como «agentes morales». Más allá de la intensa reflexión ética, jurídica, política, etc. que conlleva la IA, y que se vislumbra en una genuina preocupación social acerca de cuáles son los parámetros que han de guiar su actuación —su «programación ética»—, el debate acerca de la agencia moral no cobrará una mayor fuerza entretanto no se avance, en primer lugar, hacia una

IA general y, posteriormente, más bien, hacia una IA fuerte, capaz de funcionar como una mente, asemejada en mayor o menor medida a la humana.

VII. Referencias

o Bibliografía

- Aguiló, J. (2007). Positivismo y postpositivismo: dos paradigmas jurídicos en pocas palabras. *DOXA. Cuadernos De Filosofía Del Derecho, 30*, pp. 665–675.

- Bentham, J. (1996). *An Introduction to the Principles of Morals and Legislation.* Oxford University Press.

- Bostrom, N. (2016). *Superinteligencia: Caminos, peligros, estrategias*. Teell Editorial.

- Brozek, B. y Janik, B. (2019). Can artificial intelligences be moral agents? *New Ideas in Psychology, 54*, pp. 101-106.

- Castelfranchi, C. y Falcone, R. (2003). From Automacity to Autonomy: The Frontier of Artificial Agents. En H. Hexmoor, C. Castelfranchi y R. Falcone (eds.), *Agent Autonomy* (pp. 103-136). Springer.

- Copleston, F. (1993). *Historia de la Filosofía* (vol. VIII). *De Bentham a Russell* (2ª edición). Ariel.

- Crimmins, J. (2014). Bentham and utilitarianism in the early nineteenth century. En B. Eggleston y D. Miller (eds.), *The Cambridge Companion to Utilitarianism* (pp. 38-60). Cambridge University Press.

- Eggleston, B. (2014). Act utilitarianism. En B. Eggleston y D. Miller (eds.), *The Cambridge Companion to Utilitarianism* (pp. 125-145). Cambridge University Press.

- Julián, V. y Botti, V. (2000). Agentes Inteligentes: el siguiente paso en la Inteligencia Artificial. *Novática: Revista de la Asociación de Técnicos de Informática*, *145*, pp. 95-99.

- Honderich, T. y García Trevijano, C. (2008). *Enciclopedia Oxford de filosofía* (2ª edición). Tecnos.

- Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11*, pp. 19-29.

- Kant, I.
  (2002). *Crítica de la razón práctica*. Ediciones Sígueme.

  (1977). *Fundamentación de la Metafísica de las Costumbres* (5ª edición). Espasa-Calpe.

- Kurzweil, R. (2005). The singularity is near: when humans trascend biology. Viking.

- Lazari-Radek, K. y Singer, P. (2017). *Utilitarianism: A Very Short Introduction*. Oxford University Press.

- López de Mántaras, R. (2018). Hacia la inteligencia artificial: progresos, retos y riesgos. *Mètode: Revista de difusión de la Investigación, 99*, pp. 44-51.

- Monasterio, A. (2019). Ética para máquinas: Similitudes y diferencias entre la moral artificial y la moral humana. *Dilemata, 30*, pp. 129-147.

- Poole, D.L. y Mackworth, A. K. (2023). *Artificial Intelligence: Foundations of Computational Agents* (3ª edición). Cambridge University Press.

- Reale, G. y Antiseri, D. (1988). *Historia del Pensamiento Filosófico y Científico* (vol. II). *Del humanismo a Kant*. Herder.

- Russell, S. J. y Norvig, P. (2022). *Artificial Intelligence: A Modern Approach* (4ª edición). Pearson Education Limited.

- Searle, J.R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*(3), pp. 417-457.

- Tikhonov, A. y Yamshchikov, I.P. (2023). Post Turing: Mapping the landscape of LLM Evaluation. *ArXiv, abs/2311.02049.*

- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), pp. 433-460.

- Wallach, W., Allen, C. y Smit, I. (2008). Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI & Soc, 22*, pp. 565-582.

- Woolridge, M. y Jennings, N. R. (1995). Intelligent agents: theory and practice. *The Knowledge Engineering Review, 10*(2), pp. 115-152.

- **Otros recursos**

- Agrawal, M., Eloot, K., Mancini y M., Patel, A. (2020, 29 de julio). *Industry 4.0: Reimagining manufacturing operations after COVID-19.* McKinsey. https://www.mckinsey.com/capabilities/operations/our-insights/industry-40-reimagining-manufacturing-operations-after-covid-19

- Cameron y Garson. (2019). Connectionism. En *The Stanford Encyclopedia of Philosophy*. Recuperado el 03 de abril de 2024 de https://plato.stanford.edu/entries/connectionism/

- Márquez, J. (2023, 14 de marzo). *GPT-4 es oficial: OpenAI presenta un enorme modelo multimodal que alcanza "un rendimiento a nivel humano" en algunos escenarios.* Xataka. Recuperado el 21 de febrero de 2024. https://www.xataka.com/robotica-e-ia/gpt-4-oficial-openai-presenta-enorme-modelo-multimodal-que-alcanza-rendimiento-a-nivel-humano-algunos-escenarios

- Parlamento Europeo. (2024, 13 de marzo). *La Eurocámara aprueba una ley histórica para regular la inteligencia artificial* [comunicado de prensa] https://www.europarl.europa.eu/news/es/press-room/20240308IPR19015/la-eurocamara-aprueba-una-ley-historica-para-regular-la-inteligencia-artificial