

Estatuto ético y antropológico de las inteligencias artificiales

Ethical and anthropological status of artificial intelligence

Eduardo Gutiérrez Gutiérrez

egutierrezg@uemc.es

Universidad Europea Miguel de Cervantes

ORCID ID 0000-0001-7809-6530

Resumen: Las inteligencias artificiales forman parte de la vida cotidiana de cada vez más ciudadanos de todo el mundo: desde el estudiante que las utiliza para hacer los deberes y la anciana que ataja con ellas su soledad, hasta el empresario que implanta robots para mejorar la productividad de su fábrica o el oficinista que utiliza un chatbot para las gestiones diarias.

Que forman parte de “nuestro mundo” es un hecho indiscutible. La cuestión, por recordando el famoso libro de Scheler, es la de su puesto en ese mundo, es decir: ¿Qué son las inteligencias artificiales en el marco de “lo humano”?

En este artículo diseñamos una “teoría de teorías” para analizar la cuestión sobre el estatuto ético y antropológico de las inteligencias artificiales. Nos parece que esta es la primera piedra en el camino hacia una regulación normativa de las relaciones entre las inteligencias artificiales y las personas humanas.

Palabras clave: Cosas; dignidad; ética aplicada; inteligencia artificial; persona.

Abstract: Artificial intelligence is becoming part of everyday life for more people around the world: from students who use it to do their homework and elderly people who use it to combat loneliness, to business owners who implement robots to improve factory productivity and office workers who use chatbots for daily tasks.

That it is part of “our world” is an indisputable fact. The question, recalling Scheler's famous book, is its place in that world, that is: What is artificial intelligence in the context of “the human”?

In this article, we design a “theory of theories” to analyze the question of the ethical and anthropological status of artificial intelligence. We believe this is the first step on the road to regulatory norms governing the relationship between artificial intelligence and human beings.

Keywords: Things; dignity; applied ethics; artificial intelligence; person.

Introducción

Podemos afirmar que durante el primer cuarto del nuevo milenio las inteligencias artificiales se han convertido en un nuevo tema de debate dentro de la filosofía moral. Es este un evento inédito, pero análogo a otros acontecidos durante el siglo XX: irrupción del animalismo, ecologismo.

En efecto, el potente desarrollo que en las últimas décadas han experimentado las inteligencias artificiales suscita diversos problemas para la filosofía moral. Al efecto de clasificarlos de una manera probablemente grosera, pero funcional a los propósitos de esta introducción, utilizaremos la distinción que el profesor Vincent C. Müller establece para la entrada sobre ética de la inteligencia artificial de la famosa enciclopedia de filosofía Stanford. Distingue Müller entre “preocupaciones éticas” y “desafíos éticos”. *Preocupaciones* éticas suscitadas por la irrupción de las inteligencias artificiales son, por ejemplo, el problema para nada nuevo de la posible pérdida o

destrucción de puestos laborales, el de la seguridad digital o la manipulación de noticias e informaciones.

Las discusiones *mundanas* sobre los problemas implicados por las inteligencias artificiales caen a nuestro juicio en el grupo de las “preocupaciones ética”“, como también muchas discusiones académicas cristalizadas en libros y *papers*. Por ejemplo, el comentario a la ética de las IA que Russell y Norvig hacen en el capítulo 28 de *Artificial Intelligence: A Modern Approach* sobre las implicaciones filosóficas de estas tecnologías se reduce a la exposición, al modo de una lista de lavandería, de las preocupaciones éticas que suscitan las IA. Lo mismo sucede en “Ética y filosofía de la inteligencia artificial: debates actuales” de Santana-Soriano (2023) o “Artificial Intelligence as an Essential Tool for the Development of Medicine Requirement” (Khan et al., 2022).

Por poner un contrapunto y restarle de paso originalidad a este artículo, Mark Coeckelbergh reflexiona sobre el estatuto ético y antropológico de las IA, que denomina “político”, en el capítulo 6 de *La filosofía política de la inteligencia artificial* (2023). También se abordan estas cuestiones en el libro de Carlos Madrid Casado *Filosofía de la Inteligencia Artificial* (2024), escrito desde las mismas coordenadas filosóficas que este modesto trabajo.

Si bien no pensamos que sean irrelevantes, en este artículo queremos centrar la atención en el que probablemente sea el gran “desafío ético” que las inteligencias artificiales suscitan para la filosofía moral de nuestro tiempo: el problema de su estatuto ético y antropológico. Por explicarlo en cuatro palabras: ¿las inteligencias artificiales son *personas*? Es decir, ¿son *sujetos morales*? Si lo son, ¿en calidad de *pacientes* o de *agentes* morales?

Este desafío sirve para distinguir el problema moral de las IA de otros problemas tecnológicos, como el de la dualidad axiológica en el uso de la energía nuclear, capaz de lo mejor y lo peor para la especie humana. Hay, está claro, implicaciones éticas en el uso *técnico* que puede hacerse de la energía nuclear, pero son, por decirlo de alguna manera, de menor calado que las implicaciones éticas del uso de las inteligencias artificiales. Nadie en su sano

juicio se atreverá a discutir la condición moral o personal de la energía nuclear, pero es lícito hacerlo respecto de las IA.

La cuestión sobre el estatuto ético y antropológico de las inteligencias artificiales no sólo es central para una ética de la IA que quiera estar a la altura de los tiempos. Es, además, un verdadero desafío para la filosofía moral en general, porque nos pone ante el problema eludido por muchos de si no será necesario revisar nuestro sistema de conceptos éticos, esto es, la tradición ética tal y como la conocemos, de modo análogo a como lo propuso en 1993 el Proyecto Gran Simio de Singer y Cavalieri. Léase, por ejemplo, el artículo “En defensa de la condición de persona de los gorilas” de Francine Patterson y Wendy Gordon (1998).

A nuestro juicio, la literatura sobre la ética de la IA parece eludir las implicaciones antropológicas del problema del estatus moral de estas nuevas tecnologías. Y nos parece que, a tenor de los posibles cursos de desarrollo de las IA, que según los expertos apuntan a la aparición de una “superinteligencia” con aspecto indistinguible del humano, las implicaciones antropológicas son las más acuciantes. No se trata solamente de determinar si pueden actuar autónomamente o no. De lo que se trata, lo que tiene verdaderas repercusiones en la filosofía moral, es si debemos incluirlas o no dentro del “círculo de reconocimiento moral”.

Como siempre, a la quizás un tanto grosera afirmación que abre el párrafo anterior se pueden alegar contraejemplos como el del número especial 12 que la revista *Naturaleza y Libertad. Revista de Estudios Interdisciplinarios* dedica en el año 2019 a la relación entre la antropología filosófica y la Inteligencia Artificial, bajo el rótulo *¿Es posible transferir la mente humana a un soporte no biológico?* O también, el importante artículo “¿Es posible transferir la mente humana a un soporte no biológico?” (Linares-Peralta & Llorca Albareda, 2024), donde, en términos semejantes a los nuestros, plantean la necesidad de repensar las categorías tradicionales de la antropología filosófica.

No resulta difícil deslizarnos de la cuestión sobre el estatuto ético de las inteligencias artificiales a la de su estatuto antropológico, de manera semejante a como se hizo con los animales no humanos: a tenor de las

características humanoides que comienzan a manifestar, como la inteligencia o el aprendizaje, ¿no cabría considerar a las inteligencias humanas como personas? Si no lo son, ¿cómo justificar la aproximación ética a estas tecnologías? ¿No sería, más que un problema ético sensu stricto, un problema técnico, a saber, el de la repercusión que el uso con las tecnologías de inteligencia artificial tiene sobre las vidas de otros sujetos morales? Y si lo son, ¿en qué se diferencian de los seres humanos?

En este artículo nos proponemos un análisis filosófico del estatuto ético y antropológico de las inteligencias artificiales como prolegómeno a una ética de la inteligencia artificial. Como hipótesis de trabajo que ahora no es pertinente justificar consideramos necesario, para orientarnos por estos derroteros, partir de un determinado sistema de filosofía. En efecto, como estamos viendo en esta introducción, para responder a la pregunta central de la ética de la inteligencia artificial precisamos de una filosofía política y de una antropología filosófica, por no hablar de una filosofía de la ciencia, de una ontología y hasta de una filosofía de la mente (¿son *inteligentes* las inteligencias artificiales?, etc.). Nuestra toma de partido es por el sistema del materialismo filosófico de Gustavo Bueno, del que extraeremos el instrumental necesario para la prospección deseada.

A continuación, realizamos una breve introducción a la ética de la inteligencia artificial. Después, planteamos sucintamente los parámetros de la filosofía moral del materialismo filosófico, especialmente el artificio conceptual del “espacio antropológico”¹. En la segunda parte del trabajo localizamos el lugar que las inteligencias artificiales ocupan en el “espacio antropológico”, concluyendo con algunos comentarios sobre su estatuto ético y una síntesis de la propuesta conceptual.

¿De qué hablamos cuando hablamos de “ética de la inteligencia artificial”?

Desde la sistematización de eso que llamamos filosofía moral en la Atenas de los filósofos, el ámbito de la moralidad, que llamaremos “círculo de reconocimiento moral”, solamente protegía, dotando de derechos y *dignidad*, esto es, reconociendo como sujetos morales, al conjunto de los sujetos

operatorios humanos. Ese CRM actúa, por utilizar una metáfora luminosa, como un *paraguas* que protege del temporal de violencia y abusos que los hombres practican sobre otros hombres: a quien está bajo ese paraguas no se le puede dañar porque, por decirlo con Kant, es un fin en sí mismo y no solamente un medio; posee *dignidad* y no solo *precio*.

Es cierto que hasta hace poco tiempo el CRM solamente cubría a los seres humanos, pero no lo es que los cubría a todos sin excepción (ni siquiera hoy lo es: inmigrantes y refugiados se encuentran desprotegidos de los derechos humanos a consecuencia de la ausencia de derechos civiles y políticos). En Grecia y Roma amplias capas de la población eran expulsadas de ese círculo, como los esclavos, los extranjeros, las mujeres, los niños o los discapacitados mentales.

La declaración universal de los derechos del hombre y el ciudadano de los jacobinos revolucionarios supondría la primera propuesta positiva de ampliación del círculo a la totalidad de los seres humanos (con los antecedentes del estoicismo y de la escolástica española). A partir de este momento, el círculo de reconocimiento moral incluirá a (todas) las personas humanas bien en calidad de agentes morales (a quienes se les reconoce autonomía *moral*, es decir, responsabilidad en la toma de decisiones que afectan a otros sujetos humanos, y por consiguiente deberes u obligaciones para con los otros), bien en calidad de pacientes morales (a quienes no se les puede reconocer deberes, pero sí derechos y dignidad).

En el siglo XX, las diversas corrientes de ecoética (o *ecoéticas*) fueron ampliando progresivamente el radio del círculo de reconocimiento moral más allá de los límites antropológicos impuestos por la tradición filosófica (antropocentrismo en sus versiones fuerte y débil): sensocentrismo, biocentrismo, ecocentrismo, ética de la Tierra, etc. Durante el primer cuarto del siglo XXI el debate concerniente a lo que desde hace casi un siglo se llama “ética aplicada” gira en torno a si se deben introducir o no en ese círculo de reconocimiento moral nuevas entidades no humanas ni animales que sin embargo parecen dar muestras de la posesión de facultades y capacidades humanas: las inteligencias artificiales.

El de la ampliación del círculo de reconocimiento moral es a nuestro juicio el problema central de la ética de la inteligencia artificial (sin obviar, desde luego, la importancia de otros que en la introducción hemos llamado “preocupaciones éticas”), referente a su estatuto ético y antropológico: ¿cómo debemos considerar o concebir a las tecnologías de inteligencia artificial? ¿Nuestro trato con las inteligencias artificiales debe ser un trato ético? ¿Las inteligencias artificiales tienen derechos? ¿Y obligaciones? Este problema, aun moral, como decimos, también reverbera en la esfera antropológico-filosófica como el problema de la discontinuidad o continuidad entre seres humanos y máquinas (Linares-Peralta & Llorca Albareda, 2024, p. 4).

Problemas de la ética de la inteligencia artificial

La ética de la inteligencia artificial, como parte de una filosofía de la inteligencia artificial que por necesidad debe incluir una filosofía de las ciencias y una ontología, se ocupa de problemas muy diversos que podemos clasificar en tres grandes grupos:

Ética en el diseño de las IA

En los últimos años las tecnologías de inteligencia artificial se están aplicando a cada vez más campos: trabajo (empresa china de Shenzhen), recursos humanos (valoración de currículos para la ocupación de vacantes laborales), cuidados sociales, vigilancia y control policial, defensa y armamento militar, resolución de dudas, generación de textos e imágenes, producción de noticias, etc. Implicados a estas sucesivas aplicaciones encontramos problemas éticos como el de la manipulación o la desinformación (lindantes con la filosofía del periodismo, en el sentido del genitivo subjetivo), de los sesgos ideológicos y la discriminación, la ciberseguridad y el robo de información, entre otros. Más que problemas, hablamos de preocupaciones, de acuerdo con lo mencionado al inicio.

El problema que sirve como hilo rojo para conectar a los que hemos expuesto en el párrafo anterior es el de la *responsabilidad*: ¿a quién debemos exigir responsabilidades o rendición de cuentas? ¿A las propias tecnologías de IA? Esto nos lleva al tercer grupo de problemas que vamos a clasificar. ¿A los

inversores de las compañías productoras de esas tecnologías? ¿A los usuarios?, ¿a sus entrenadores o diseñadores?

Este es el principal foco de problemas que se está discutiendo hoy en día, fundamentalmente desde el Derecho, estableciendo límites legales en el uso y diseño de las IA en base a principios de transparencia, seguridad, justicia y no discriminación, respeto de la autonomía humana, responsabilidad y rendición de cuentas. Digamos, por consiguiente, que este primer grupo de problemas se encuentra a medio camino entre la ética de las inteligencias artificiales y el Derecho: (Cerrillo, Martínez & Peguera Poch, 2020), (Garapon & Lassègue, 2021), (Herrera Triguero, Peralta Gutiérrez & Torres López, 2022).

Ética de las IA (o para las IA)

Toda vez que reconocemos el grado de autonomía que las IA de última tecnología están adquiriendo para la toma de decisiones con poderosas implicaciones antropológicas, ¿no sería conveniente introducirles algorítmicamente sistemas éticos que regulen esa toma de decisiones?

Según explican Anderson y Anderson (2007), la ética de las máquinas refiere a un ámbito de análisis filosófico moral que reflexiona sobre el comportamiento que las IA, o las máquinas en general, tienen hacia sus usuarios humanos. Por ejemplo, el comportamiento que un robot programado para cuidar de un anciano en una residencia tiene hacia este.

El ejemplo que suele utilizarse como hipótesis para este tipo de discusiones es el del coche autónomo que se enfrenta a una de esas situaciones que llamamos “dilemas éticos”, en la que se plantean diversos cursos de acción entre los que es preciso elegir justificada o razonablemente: ¿cuál es la decisión que *debería* tomar la máquina? En la novela de Asimov *I, robot* esto se ve claramente: si una máquina debe decidir entre salvar la vida de una niña pequeña y la de un hombre adulto de mediana edad, ¿a quién elegiría salvar y por qué? Y lo que es más importante: ¿Podría justificar su toma de decisión aportando argumentos moralmente significativos (válidos para todos)?

El problema anterior no parece revestir dificultades reseñables. Mayoritariamente se convendrá en la necesidad de introducir un sistema ético que facilite la toma de decisiones de las máquinas en favor de los humanos. Sin embargo, ¿cuál de los muchos sistemas éticos realmente existentes en la tradición filosófica, y entre los cuales hay incompatibilidades dialécticas, es el apropiado para este diseño normativo?, ¿quién tiene la autoridad para decidir? ¿Es mejor el deontologismo o el utilitarismo? ¿Ayudan estos sistemas de la tradición a tomar decisiones en un mundo como el nuestro? A juicio de Brožek y Janik (2019), es más fácil para una máquina que para un ser humano atenerse a las máximas de sistemas éticos tan abstractos como el deontologismo kantiano o el utilitarismo de Bentham. Esta paradoja dice poco en favor de esos sistemas morales, cuya idealización los aleja completamente de la práctica moral realmente existente.

El primer sistema de normas éticas (no ya jurídicas) para el diseño de las Inteligencias Artificiales lo constituyen las tres famosas leyes de la robótica de Asimov (*I, Robot*, 1950): Un robot nunca hará daño a un ser humano ni permitirá que un ser humano sufra daños; Un robot deberá obedecer las órdenes que le dé un ser humano, salvo en los casos en que tal obediencia contravenga la primera ley; Un robot deberá proteger su propia existencia, siempre que dicha protección no entre en conflicto con las otras dos leyes.

Ahora bien, es discutible que un sistema de normas que se implantan en un sujeto operatorio inanimado durante su proceso de diseño matemático e informático para que actúe en conformidad a ellas, como si se tratara de pautas de conducta más que de normas en el sentido estricto del concepto, pueda ser llamado un sistema ético.

Para que podamos hablar de sistema ético, o de ética en sentido amplio, se requiere de la autonomía del agente moral como condición necesaria. A grandes rasgos, distinguimos dos tipos de autonomía: autonomía cognitiva (conocimientos para realizar alguna operación compleja de manera independiente, sin asistencia) y autonomía moral (responsabilidad, conciencia de los motivos y las consecuencias de nuestras acciones). Un ejemplo para captar la diferencia entre ambos tipos: un individuo al que se le expide el carné

de conducir ha demostrado poseer autonomía en el sentido cognitivo para la práctica de conducción de un automóvil. Sin embargo, ese mismo individuo, por descuido, puede atropellar a una persona que cruza un paso de cebra y demostrar falta de autonomía moral si decide no asistirle y huir, no haciéndose con ello cargo de las consecuencias de su acción.

Si no hay autonomía (en el sentido moral), no hay ética más que como cuidado hacia el paciente moral. En todo caso, no tenemos agente moral que actúa conforme a normas. Así las cosas, parece que las inteligencias artificiales no son *autónomas* sino *automáticas*, y las normas en virtud de las cuales se las entrena no son normas éticas sino comandos de acción semejantes a los que el usuario de videojuegos transmite a su avatar digital.

En este grupo de problemas cabe incluir la reflexión sobre la incidencia que una extensión en el uso con las IA puede tener sobre las relaciones humanas, en términos de pérdida de habilidades cognitivas (Greenemeier, 2011) y personales (Groth, Nitzberg & Esposito, 2018), o de erosión de la sociedad civil (como se demostró en el año 2018 con el escándalo Cambridge Analytica-Facebook).

Ética con las IA: la cuestión de la “agencia moral”

En nuestro día a día *tratamos*, ya no simplemente *utilizamos*, con las inteligencias artificiales. La batería de problemas clasificados en los grupos anteriores advierte sobre los peligros que este trato puede entrañar para los seres humanos, mientras este tercer grupo nos pone ante el espejo de nuestras propias acciones y conductas *morales*: ¿debemos considerar a las IA como sujetos morales? Es decir, ¿debemos regular normativamente nuestro trato con las IA, como un trato moral con entidades personales?

Según informan algunos diarios haciéndose eco de recientes y divertidos memes, hay un porcentaje considerable de usuarios que tratan cortés o educadamente a las IA con las que interactúan diariamente (chatbot generalmente) por miedo a que se vuelvan hostiles y peligrosas en el futuro.

Esto remite al argumento del antropologismo débil defendido por filósofos como Santo Tomás o Kant: hay que regular el trato entre los seres humanos y los seres no humanos, no por el valor intrínseco que les reconocemos a los animales sino por su valor *para sí*, para la especie humana: el ejercicio de la violencia contra un animal podría llevarnos por la pendiente resbaladiza de la violencia hacia otros humanos. *Mutatis mutandis*, en este caso tratamos bien o moralmente (educadamente) a las máquinas no porque las consideremos sujetos morales, sino por miedo a que el día de mañana se rebelen contra nosotros.

La urgencia de este problema puede ponerse en comparación con el despertar de la conciencia animalista. Hasta el siglo XIX se pensaba que entre los seres humanos y los animales existía un abismo teológico, perfectamente fundamentado por el mecanicismo de raigambre cartesiana: los animales son pura *res extensa*, máquinas que se mueven instintivamente a la manera de autómatas, mientras los seres humanos son *res pensante*, entidades con espíritu. La causa de ese *gorismós* es la insuflación del alma en el cuerpo animal de los hombres por la vía divina.

La teoría darwiniana de la evolución fue el punto de inflexión que comenzó a tambalear el dogma mecanicista, aduciendo una continuidad biológica entre animales y humanos. La genética moderna vendría a confirmar la ruptura del hiato y a demostrar una semejanza inaudita entre los humanos y nuestros parientes zoológicos.

Huelga decir que la filiación biológica, genética o fenotípica entre los humanos y las tecnológicas de IA es imposible. Sin embargo, hay una analogía entre ambos casos que se puede explicar por vía de la intuición empírica. Antes de Darwin y de las teorías predecesoras del darwinismo, se intuía una semejanza conductual entre los animales y los humanos. Así, por ejemplo, el sensocentrismo: si los animales son capaces de sentir, ¿por qué no incluirles como miembros de pleno derecho en el círculo de reconocimiento moral? Dicha intuición empírica se puede aplicar perfectamente al caso de las inteligencias artificiales, que no en vano son consideradas efectivamente como inteligentes (otro tema que no nos ocupa aquí es el de si de hecho lo

son, y qué es la inteligencia). Si las IA razonan, aprenden y piensan, como lo hacemos los humanos, ¿por qué no incluirlas también en el círculo de reconocimiento moral?

En caso de aceptar esa inclusión, y de reconocer que las IA son sujetos morales, ¿en calidad de pacientes o de agentes morales? Son autónomos y toman decisiones, luego parece que son más agentes que pacientes morales.

Para salvar la pendiente resbaladiza por la que amenaza conducirnos esa última declaración se ha acuñado la categoría de “agentes morales funcionales”, referente a sujetos morales que pueden tomar decisiones significativas, es decir, con consecuencias que afectan a la vida de las personas humanas, pero sin intencionalidad. Van Wynsberghe y Robbins (2019) utilizan el concepto de “agentes morales autónomos”. Otros, como Allen, Varner y Zinser (2000), prefieren el de “agentes morales artificiales” para limitar las expectativas de su compromiso.

Mandy Zafar (2024) niega que las IA sean agentes morales, y propone el concepto de “AI automated performance” (AIAP). Viene a decir que en las discusiones recientes se utiliza el concepto de agente moral de una manera demasiado laxa y simplista, atribuyéndoles a las tecnologías de IA tal estatus moral en virtud de su autonomía, racionalidad e inteligencia. El concepto de agente moral se utiliza preferentemente para los seres humanos, aunque por analogía lo empleemos también para interpretar la conducta de otros seres no humanos como nuestras mascotas. Dice a este respecto:

Llamar “agentes” a los animales o decir que son capaces de tener acciones voluntarias no significa automáticamente que sean agentes morales. Ser considerado un “agente moral” conlleva repercusiones normativas adicionales, entre ellas la atribución de capacidades jurídicas o la recepción de derechos legales. Sin embargo, en algunos casos se atribuyen derechos legales sin una clara atribución de agencia moral, como ocurre, por ejemplo, con los animales. Existe un amplio consenso en que los animales son al menos receptores de bienestar, pero no actores de voluntad propia. Esta es una cuestión abierta al

debate. Aunque estos temas son muy interesantes e importantes, no se pueden tratar con más detalle aquí. (Zafar, 2024, p. 2)

James Moor (2006) clasifica cuatro tipos de agentes morales: agentes de impacto ético (un conductor robot), agente ético implícito (un piloto automático), agente ético explícito (capaz de realizar juicios en base a la utilidad de cada alternativa) y agente ético completo (que hace juicios éticos y es capaz de justificarlos). La pregunta es: ¿Se pueden crear agentes éticos completos, categoría hasta ahora solamente reservada para los seres humanos, *artificialmente*?

Además, puestos a utilizar la idea de autonomía como criterio para la concesión de agencia moral, convendría distinguir dos concepciones: una cosa es la autonomía como saber hacer algo, o autonomía cognitiva, y otra la autonomía en el sentido moral, ligada a la responsabilidad. Cuando a alguien se le expide el carné de conducir es porque ha demostrado ser capaz de conducir autónomamente sin necesidad de un profesor acompañante en el asiento del copiloto. Decimos que ese conductor es autónomo o tiene la capacidad de conducir autónomamente. Sin embargo, si ese conductor una noche va despistado y por saltarse un paso de cebra atropella a una persona y huye, incurriendo en un delito de omisión de socorro, demuestra no tener autonomía moral, esto es, responsabilidad, capacidad para hacerse cargo de las consecuencias de sus acciones.

El problema de la autonomía de las inteligencias artificiales, tal y como lo estamos exponiendo, está directamente relacionado con el de su racionalidad.

Pero a nuestro juicio el de la racionalidad no es un criterio suficiente para la inclusión de un sujeto en el círculo de reconocimiento moral. La etología ha demostrado que todas las especies animales son racionales (“raciomorfas”, se dijo durante una época para marcar las diferencias), pero no es la racionalidad lo que las incluye en dicho círculo; de hecho, muchas de ellas ni siquiera son incorporadas, como los insectos. Ninguna de las ecoéticas que reivindican la incursión de las especies animales en el círculo de reconocimiento moral apelan a la racionalidad como criterio justificativo. Es

más, la racionalidad fue el criterio que en tiempos sirvió para dejar fuera del círculo de reconocimiento moral a discapacitados mentales, niños, ancianos dementes y extranjeros (sin *logos*).

En su conocidísimo manual *Artificial Intelligence: a modern Approach* Russell y Norvig ofrecen una definición de la acción racional que a nuestro juicio resulta insatisfactoria: “Un agente racional es aquel que actúa de manera que logre el mejor resultado o, cuando hay incertidumbre, el mejor resultado esperado” (2026, p. 22).

Esa distinción, que ejerce una idea de racionalidad teleológica, no discrimina los *finis operantis* a los que se propone llegar el agente racional de los *finis operis* a los que la acción llega objetivamente, con independencia de que el agente racional buscase o no tal fin. Lo que nos interesa en esta ocasión es la preparación proléptica del *finis operantis*, porque en la ejecución de la acción pueden influir diversos factores capaces de alterar su curso. La estrategia racional es, en efecto, proléptica, es decir, proyectada hacia el futuro, pero en toda prolepsis se implica una anamnesis o recuerdo en el que se arrastran normas institucionalizadas (y socializadas) que contribuyen a la preparación de la estrategia. Por lo tanto, la racionalidad, más que en las ideas internas del sujeto racional, está en las instituciones y las normas segregadas de estas en las que ese se apoya para el cumplimiento de *sus finis operantis*. A juicio de Bostrom y Yudkowsky (2014), una inteligencia artificial no tiene propósitos o fines en el sentido de *finis operantis*. Han sido sus programadores quienes las crean en aras a dichos objetivos, en el sentido de *finis operis*.

Como los propios autores comentan en el capítulo 28 del libro, el diseño computacionalista de las IA, lo que se llama Good Old-Fashioned AI, es absolutamente incapaz de replicar esa dimensión normativo-institucional fundamental para explicar la racionalidad humana. Es lo mismo que defiende el “argumento de la informalidad del comportamiento” de Turing: “el comportamiento humano es demasiado complejo como para ser capturado por un conjunto formal de reglas [lógicas, binarias, simples]” (Russell y Norvig, 2023, p. 1033). Los sistemas de razonamiento probabilístico y el *deep*

learning podrían, a su juicio, lograr la replicación más o menos fiel de una racionalidad compleja no sujeta a normas lógicas o silogísticas.

En el siguiente apartado esclarecemos cuál es nuestra definición de la “agencia” o el “agente” moral.

Por último, también cabe analizar los problemas éticos derivados del papel que las IA pudieran desempeñar en la vida buena de los seres humanos, que según Sócrates es la pregunta fundamental de la ética. Sin embargo, el de la “vida buena” o “lo mejor” es un criterio muy poco apto para las reflexiones concernientes a la ética de la inteligencia artificial porque no queda del todo claro qué vida es la buena ni para quién lo es. Además, las virtualidades para la vida buena que las IA ofrecen por ejemplo a los clientes de una sucursal bancaria suiza que ha implantado esta tecnología para agilizar y dinamizar su servicio de atención al cliente pueden ser contraproducentes o irrelevantes para un solicitante de hipoteca de raza negra (según el ejemplo con el que abren el capítulo 14 de su libro Bostrom & Yudkowsky, 2014).

Filosofía moral del materialismo filosófico

En este apartado hacemos una breve reexposición de la filosofía moral del materialismo filosófico, basándonos en el libro de 1996 *El sentido de la vida*.

Moralidad y normas

Definimos la “moralidad”, en un sentido amplio, como el conjunto de las normas (éticas, morales y jurídicas) que regulan las relaciones interpersonales; es decir, las relaciones mantenidas entre sujetos operatorios humanos a quienes se les reconoce la condición ética y jurídica de personas.

La idea de persona tiene como contexto genético-operatorio la escena teatral griega, concretamente, la máscara que los actores portaban para salir a escena y representar a su personaje (prosopón, per sonare). En base a este origen tecnológico, concebimos la categoría de persona más como una institución sociohistórica que como un atributo inherente a los seres humanos:

Persona humana añade algo no sólo a “persona” sino también a “humano”. *El hombre recibe una determinación importante cuando se le considera como persona así como la persona recibe una determinación no menos importante cuando se la considera como humana.* Por tanto, no es lo mismo hombre que persona, como tampoco es lo mismo hombre que ciudadano. (Bueno, 1996, p. 115)

La persona es una institución con una historia que se abre paso en sociedades desarrolladas con un lenguaje articulado de pronombres personales y con leyes que protegen la propiedad privada (de familias, más que de individuos). Sólo dadas estas condiciones es posible la operación jurídica de reconocimiento de la condición de persona a un sujeto operatorio, siempre en el marco de una comunidad de personas. Y tal como una persona recibe ese atributo histórico por vía institucional, puede perfectamente serle retirado (considerándole “persona cero”). Como no podemos extendernos más en este punto, remitimos al artículo de Gustavo Bueno “Para una construcción de la Idea de Persona” (1953).

La idea central de la filosofía moral del materialismo filosófico es la de Norma (en un sentido antropológico, más que jurídico). Gnoseológicamente, las normas desempeñan en el ámbito de las ciencias humanas un papel análogo al de las leyes naturales en las ciencias naturales. La diferencia es que las leyes normativas pueden ser justificadas y desobedecidas, mientras las leyes naturales pueden ser explicadas y no desobedecidas, pero sí controladas.

Los códigos normativos son conjuntos de instrucciones que han cristalizado al calor de ciertas instituciones antropológicas y regulan la ejecución de procesos operatorios realizados sobre entidades humanas o no humanas prolépticamente, esto es, con vistas a la consecución de determinados fines referidos a la supervivencia y progreso de un grupo social. La consolidación de una norma para la praxis humana se produce cuando en la confluencia de diferentes rutinas operatorias habilitadas para configurar objetos o situaciones repetibles, como la caza o el fuego, unas prevalecen sobre otras que resultan vencidas aunque se mantienen como virtualmente

posibles. Las normas, por decirlo en cuatro palabras, son rutinas victoriosas. Además, nos sirven como criterio para distinguir, rompiendo con la distinción clásica, la *ética* de la *moral*.

Ética y moral

El materialismo filosófico se opone a la interpretación tradicional (presente en todos los manuales de ética) entre la ética y la moral, según la cual la moral es un sistema normativo que rige para una comunidad y la ética la reflexión de segundo grado sobre esa praxis moral (filosofía moral o *meta-moral*). La famosa distinción escolástica entre la *ethica utens* y la *ethica docens*.

Dicha distinción está fundada en una oposición falsa, que maneja ideas oscuras y confusas, entre teoría y práctica. En la práctica de un mecánico hay tanta teoría como en los apuntes de un estudiante de ingeniería, y en las teorías científicas están implicadas las prácticas operatorias (con *aparatos*) de los experimentos. En resolución, en la praxis moral hay teoría o conocimientos sobre las acciones morales, lo mismo que en la teoría moral, por especulativa que sea, hay implicaciones prácticas.

Además, Bueno intuye un interés gremial por justificar la necesidad de un cuerpo de expertos en ética cuya función es la de enseñar a los alumnos a pensar éticamente, como si no supieran hacerlo ya de facto. La filosofía moral académica, de los filósofos y los profesores de filosofía, procede *in medias res* de la filosofía moral mundana o el conocimiento teórico y práctico que toda persona posee acerca de la moralidad en el sentido amplio.

En efecto, sobre la “moralidad” en el sentido amplio antes comentado hay sin duda un saber mundano, una “moralidad mundana” o “sabiduría moral mundana” que incluye una práctica moral y un conocimiento sobre esa práctica moral. Como afirmó Sócrates por pluma de Platón, todos los ciudadanos poseemos ciertas nociones sobre la praxis moral en virtud de las cuales tenemos la capacidad para emitir juicios en los que valoramos las conductas y acciones propias y ajenas. Si un individuo observa un ataque violento contra un paciente moral (niño, mujer, anciano, hombre, perro... ¿robot?) sentirá rabia, enfado o hasta tristeza, y juzgará que tal acción

interpersonal es reprobable. “No se debe actuar de este modo”, dirá, queriendo con Kant que su máxima se eleve a la condición de ley universal. Probablemente mediará, aun poniendo en riesgo su propia salud física, para detener el ataque. Y eso, pensarán otros, es una acción virtuosa.

Así, por traer esta cuestión a nuestro objeto de análisis, cuando un espectador anónimo (que no es profesor de filosofía ni ha recibido formación filosófica más allá de la formación básica de la enseñanza media) ve en vivo y en directo o a través de las pantallas a un individuo o a un grupo de individuos golpear deliberadamente (quizás para evaluar su respuesta, en una suerte de experimento científico) a un robot, experimentará esas emociones que son indicio psicológico de una ofensa moral y podrá en todo caso emitir un juicio sobre el valor moral de dicha acción.

Si preguntamos a un grupo de estudiantes de bachillerato, de Universidad, o a una selección al azar de profesionales de diversos ámbitos, tendrán, sin duda, una opinión bien formada acerca del estatus moral y antropológico de las IA. Estas consideraciones, por prosaicas que sean, deben ser recogidas y analizadas críticamente por el filósofo profesional o académico.

Este saber moral mundano es semejante al saber práctico del que hablaba Aristóteles, la prudencia o *prhonesis*, distinto del saber especulativo o ciencia (*episteme*). Incluye la *sindéresis*, es decir, el conocimiento de los primeros principios prácticos. Aunque como apunta Bueno contra Kant el saber moral mundano no es universal sino plural, referido a la diversidad de pueblos realmente existentes, cabe la posibilidad de reconocer un puñado de principios morales que si no son de hecho universales podrían considerarse de tal naturaleza al menos virtualmente.

Por ejemplo, la regla de oro (no hagas para los demás lo que no quieras para ti), o, por lo menos en la tradición de la ética occidental contemporánea, los cuatro principios de la bioética (autonomía, beneficencia, no maleficencia, justicia). Efectivamente, estos principios sirven para regular la praxis profesional, que no deja de ser un caso especial de praxis moral entre sujetos operatorios humanos cuyas consecuencias implican al conjunto de la

sociedad (del mismo modo que la conducta de *bullying* de un grupo de jóvenes contra otros a través de las redes sociales tiene repercusiones sociales de gravedad).

Claro que lo dicho no es óbice para la existencia de “idiotas morales” (utilizando el vocablo en su acepción griega, *idiotés*, el que va a la suya). El idiota moral es quien habiendo cometido un acto horrendo se siente orgulloso de ello, no arrepintiéndose aun teniendo conciencia de lo que ha hecho. Tampoco es óbice para la existencia de casos patológicos de incapacidad para adquirir conciencia de culpa de los actos propios, como la psicopatía.

Lo mismo que existe un saber moral mundano, existe un saber moral académico o “moralidad académica” en relación dialéctica con este, y del que bebe como una de sus fuentes básicas. Es lo que llamamos comúnmente filosofía moral (en el sentido estricto o académico). La filosofía moral no tiene la misión de instaurar la conciencia moral, porque su misma existencia ya presupone el juicio moral. Pero de aquí no se deduce que la filosofía moral sea superflua, prescindible y mero saber de especialistas sin utilidad política alguna. No enseña a pensar éticamente, sino a discriminar entre las diversas alternativas en juego en los muchos dilemas éticos a los que nos enfrentamos diariamente.

Veamos desde esta luz y a partir del criterio de las normas la distinción que propone Bueno entre la ética y la moral.

El criterio que se utiliza habitualmente para distinguir ética y moral es el del origen de las normas que regulan la conducta interpersonal de los sujetos humanos: la ética tendría su punto de partida en la conciencia o intimidad de la persona humana, y la moral en la sociedad o el colectivo. Sin embargo, la crítica al reduccionismo subjetivista sería suficiente para negar que en la conciencia del individuo puedan brotar nociones o concepciones éticas con absoluta independencia del contexto social de aprendizaje y formación intelectual. Por nuestra parte, el criterio que utilizamos es el del *finis operantis* de las normas; es decir, no su punto de origen sino de llegada, su *término*.

La ética y la moral son sistemas normativos para regular la praxis interpersonal, solo que a diferente escala. El punto común de ambos sistemas normativos es su objetivo o fin: la preservación de los cuerpos operatorios de los sujetos personales, sin los cuales no hay, frente al espiritualismo, praxis ética alguna. La *sindéresis* o principio práctico conocido por todos, la ley fundamental de la conducta ética y moral, dice así:

Debo obrar de tal modo que mis acciones puedan contribuir a la preservación en la existencia de todos los sujetos humanos, y yo entre ellos, en cuanto son sujetos actuantes, que no se oponen, con sus acciones u operaciones, a esa misma preservación de la comunidad de sujetos humanos. (Bueno, 1996, p. 57)

Este principio fundamental se divide en dos leyes correspondientes a los contextos atributivo y distributivo en los que se da la existencia de los sujetos corpóreos: en el contexto atributivo (la humanidad descompuesta en grupos, tribus, bandas, clases o Estados) el principio obliga a actuar en aras a la preservación de la comunidad; en el contexto distributivo (la humanidad descompuesta en sus unidades mínimas -que no básicas-, los individuos), obliga a actuar en aras a la preservación de los individuos corpóreos. La modulación atributiva de la *sindéresis* establece las normas morales, ordenadas a garantizar la supervivencia o recurrencia en el tiempo de la comunidad. La modulación distributiva establece las normas éticas, referidas a la preservación en la existencia de los sujetos individuales corpóreos. La ética es a la medicina (salud del cuerpo) lo que la moral es a la política (salud social).

En el ámbito trascendental (respecto de la conciencia individual) de las normas éticas la virtud fundamental es la fortaleza o fuerza del alma de Baruch Espinosa, que se manifiesta como firmeza cuando el esfuerzo se dirige a la conservación del propio ser (cuerpo operatorio) y como generosidad cuando se orienta a la ayuda a los demás.

En el ámbito también trascendental (no relativista) de las normas morales no hay aplicación posible de la firmeza. Su equivalente será el grado de cohesión del grupo. Lo mismo cabe decir de la generosidad, que con

respecto al grupo se aplicará sobre otros no por generosidad ética sino por cálculo político, a no ser que se las relaciones de ayuda se interpreten retóricamente como relaciones de generosidad. Hablamos, no de generosidad, sino de solidaridad (contra terceros). El principio fundamental de la moral es la justicia, entendida en el sentido de seguimiento escrupuloso de las normas que regulan las relaciones de los individuos en cuanto partes del todo social.

Entre las normas morales y las normas éticas puede haber relaciones de compatibilidad, cuando para la aplicación de la justicia se requiere de la aplicación de la generosidad o de la firmeza (la cohesión de mi grupo no puede mantenerse si no soy generoso con aquellos sujetos operatorios que lo integran). Es más, las normas éticas comienzan a abrirse paso en el contexto de la familia. Pero también pueden darse relaciones de incompatibilidad, cuando para garantizar la cohesión o recurrencia de la comunidad es preciso retirar la generosidad hacia ciertos sujetos operatorios, eliminarlos de un puesto social (haciendo que su vida peligre, al perder la base económica de energía vital), o hasta neutralizarlos (reconociéndolos como “personas cero”). Hay varios ejemplos de conflicto entre las normas éticas y las normas morales: ejecución capital de ciertas categorías homicidas o criminales, institución italiana de la *vendetta*, casos históricos de sacrificios humanos y hasta de infanticidio, dilema del soldado caído, etc. Por decirlo en pocas palabras, hay situaciones límite (pero posibles) en las que para que conservar el horizonte de normalidad que mantiene a la comunidad unida y orientada prolepticamente hacia el futuro es preciso neutralizar a ciertos sujetos operatorios a quienes se les habría retirado, por razones muy diversas (comisión de actos atroces, discapacidad, ausencia de capacidad operatoria, entorpecimiento, etc.), su condición de persona.

Teoría del espacio antropológico

En este subapartado damos un salto de la filosofía moral (y política) a la antropología filosófica. Un salto importante, desde luego, pero justificado por el enfoque de nuestro artículo. Porque, con arreglo a lo que se ha explicado arriba, para que las inteligencias artificiales posean un estatus ético es

condición necesaria que se les pueda reconocer un estatuto antropológico previo: si no son personas o proyecto de persona, no hay hacia ellas obligaciones éticas algunas. Otra cosa son las obligaciones morales que podamos tener hacia dichas tecnologías como recurso para garantizar la supervivencia del grupo, análogas a las que una comunidad tiene respecto de los elementos impersonales de su entorno natural en aras a lograr el “equilibrio ecológico” de Marvin Harris.

Para determinar el estatuto antropológico de las inteligencias artificiales remitimos a la teoría del espacio antropológico del materialismo filosófico, a partir del artículo “Sobre el concepto de ‘espacio antropológico’” (Bueno, 1978).

La idea del “espacio antropológico” sustituye a la problemática idea de Humanidad cuanto que *objeto* de estudio de la Antropología positiva (“ciencia del ser humano”). Es crítica de la concepción sustancialista de la Humanidad como esencia eterna dada en un momento dado desde el que ha mantenido una relativa invariancia. Los propios estudios de antropólogos, etólogos y zoólogos demuestran que esa supuesta esencia invariante es resultado de un largo proceso de desarrollo evolutivo (biológico y cultural) en el que actúan determinadamente los contextos ecológicos envolventes y las realidades no antropológicas dispuestas en ese escenario natural. Por lo que la idea de Humanidad debe incluir necesariamente esos materiales no humanos.

Esta es la primera consideración sobre la que se levanta nuestra teoría del espacio antropológico: el campo, que no el objeto, de la Antropología (y de la filosofía moral), no es propiamente el Hombre sino lo que llamamos “material antropológico”, distribuido en tres clases: (i) la clase de las personas, (ii) la clase de las cosas culturales (que median en las relaciones interpersonales: armas, instrumentos, documentos, tecnologías, etc.), y (iii) la clase de las acciones y de las operaciones (en el ámbito moral, donaciones, omisiones a injusticias, salvaciones, mentiras, agresiones, cooperación, generosidad, etc.). Estos materiales antropológicos no están exentos, como

pensaría un idealista al estilo de Fichte, sino envueltos por un espacio antropológico en el también hay materiales no antropológicos o no humanos.

En realidad, esta idea del espacio antropológico está presente en la tradición filosófica, si bien más ejercida que representada. Está en Descartes, por ejemplo, cuando clasifica los entes del mundo en *res pensante* o espíritu y *res extensa* o cuerpo. Está también en Kant, cuando distingue los fenómenos según sean Cultura o Naturaleza, y en Hegel y Marx, quienes transforman los términos kantianos para hablar, como Fichte, de Espíritu o Cultura (Yo) y Naturaleza (No-Yo).

El espacio antropológico del materialismo filosófico, pluralista, niega el dualismo metafísico en el que han incurrido los ejercicios anteriores de la idea de espacio antropológico. Como resultado de esa negación dialéctica (es decir, con la potencia para incorporar lo negado), el espacio antropológico del materialismo filosófico no tiene dos sino tres ejes, porque tres son las ideas centrales de la ontología clásica desde Aristóteles: Hombre, Mundo y Dios. Solo que poniendo el principio del ateísmo esencial el materialismo filosófico niega la idea de Dios, estableciendo la referencia real y objetiva de la religión en los *númenes*, animales (de la megafauna del Pleistoceno) que no siendo humanos se parecen a los humanos y son percibidos por estos como fuerzas sobrenaturales de las que depende su vida y su muerte.

Los tres ejes del espacio antropológico son:

- (a) Eje circular de las relaciones entre los entes personales e inmanentes, es decir, entre las personas humanas: relaciones jurídicas, sociológicas, económicas, políticas, familiares, etc.

Las relaciones éticas y morales en sentido estricto son relaciones circulares. Si las IA están en el eje circular es porque se les reconoce un estatuto personal: son personas o proyecto de persona, y, en consecuencia, las relaciones que los seres humanos mantienen con ellas son relaciones morales como las mantenidas entre los propios seres humanos. El estatuto personal implica el estatuto ético (poseen derechos).

- (b) Eje radial de las relaciones entre los entes personales e inmanentes y los entes impersonales y trascendentes, es decir, entre las personas humanas y las cosas inanimadas e impersonales (artefactos, instrumentos, aparatos, tecnologías, entidades naturales, etc.): relaciones de producción agrícola y fabril, minería, siderurgia, industria, arte, etc.

No es posible, más que en los delirios del enfermo mental, mantener relaciones éticas (de firmeza y generosidad) con entes inanimados e impersonales. La relación entre el sujeto operatorio humano y el ordenador es una relación de carácter técnico, pero no ético. Y aunque la idea de Bien está involucrada en las relaciones tecnológicas, no lo está en la misma medida en que interviene en las relaciones éticas. Un ejemplo para explicar esto.

Una cámara de gas es una tecnología *buena* desde el punto de vista técnico, porque sirve con sobrada eficacia para el objetivo para el que fue diseñada por los nazis: la ejecución rápida y *limpia* de la mayor cantidad de prisioneros judíos posible. Es *buena*, como *bueno* es el látigo que lacera y arranca la carne de las espaldas del nazareno. Buena técnicamente hablando, como es *bueno* mi teléfono móvil porque me ofrece la posibilidad de comunicarme con mis seres queridos a miles de kilómetros de distancia. “Bueno”, aquí, en el sentido de utilidad técnica, “bien técnico” o instrumental. Sin embargo, la cámara de gas es un artefacto “malo” o “vicioso” (por ausencia de generosidad) desde el punto de vista ético, porque supuso la neutralización de la vida de cientos de miles y hasta de millones de seres humanos, esto es, de sujetos con derecho y dignidad.

Las relaciones entre los sujetos operatorios humanos y los animales destinados al consumo humano, como las que se producen en un matadero, son relaciones radiales. El animal no es visto por el matarife como una entidad subjetual, esto es, como una persona (eje circular), ni como un numen divino (eje angular): es una cosa, un medio de subsistencia. *Emic*, desde la perspectiva del matarife, no hay infracción a norma ética alguna en la aplicación de la muerte del animal; si acaso, en los modos de aplicar la muerte. Es más, en una situación de escasez de recursos que pone en riesgo de supervivencia a la

comunidad el sacrificio del animal para alimentar a la población es una obligación moral, sean cuales sean las consideraciones ideológicas del matarife (que bien podría ser animalista o sensocentrista). La misma lógica de incompatibilidad se aplica al caso del canibalismo en situaciones de urgencia vital.

- (c) Eje angular de las relaciones entre los entes personales immanentes y los entes personales trascendentes, es decir, entre las personas humanas y las personas no humanas (animales numinosos, ángeles, demonios, extraterrestres, dioses apotéticos). Las relaciones angulares son relaciones de carácter religioso, a tenor de la filosofía de la religión del materialismo filosófico.

Difícilmente las relaciones angulares o religiosas pueden ser calificadas de relaciones éticas, más allá de las demandas éticas que se segregan de las doctrinas religiosas. Pero aquí no hablamos propiamente de religión en el sentido doctrinal, esto es, del cristianismo, judaísmo, islamismo, etc., sino de las relaciones *religiosas* entre humanos y entidades numinosas (religión como derivado etimológico de *religatio*, religación, relaciones operatorias).

Este tipo de relaciones se mantienen, en la terminología nietzscheana, más allá del bien y del mal. ¿Cómo aplicar el imperativo categórico con una entidad numinosa con la que es imposible la reciprocidad operatoria, de la que ni siquiera tengo una referencia apotética? Según Aristóteles, Dios no crea ni conoce el mundo, por lo que tampoco profesa sentimientos de índole religiosa hacia los humanos.

Emic, el homínido prehistórico puede pensar que el mamut que decide no investir al grupo de cazadores y cambiar el rumbo de su marcha está actuando misericordiosa o incluso generosamente, perdonándoles la vida. Sin embargo, no es posible demostrar que los animales (en estado de salvajismo) actúan motivados por disposiciones éticas (más fácil podría ser la atribución analógica de disposiciones morales, orientadas a la supervivencia del grupo).

La “agencia moral” según el materialismo filosófico

De acuerdo con los parámetros sistemáticos que acabamos de reexponer, la agencia moral es un concepto de naturaleza ética (más que moral) que se dibuja siempre en el eje circular del espacio antropológico: solamente las personas humanas son sujetos de quienes quepa predicar una agencia moral o capacidad para obrar moralmente; por consiguiente, los agentes morales son aquellos sujetos humanos corpóreos a quienes cabe exigir el cumplimiento de las normas éticas de la generosidad y la firmeza. Los pacientes morales son aquellos sujetos humanos corpóreos a quienes no se les puede exigir la norma de la generosidad, en quienes se presupone el seguimiento de la firmeza (solo que de una forma más bien instintiva, espontánea, a la manera del *conatus* espinosista o de la *naturaleza* estoica), y sobre quienes los agentes morales tienen la obligación ética de la generosidad.

Como ya hemos sugerido más arriba, ni la autonomía ni la racionalidad son criterios suficientes para la concesión de agencia moral a los sujetos corpóreos, por sus implicaciones metafísicas. Más pertinente es el concepto de *responsabilidad*, coordinable con nuestra norma moral de la *prudencia* (orientada a la supervivencia del grupo, o, en términos más bien políticos, de la nación: *eutaxia*).

Las inteligencias artificiales en el espacio antropológico

Al barruntar cuál es el estatuto antropológico, y, consiguientemente, ético de las inteligencias artificiales desde las bases del materialismo filosófico, nuestro objetivo es determinar en cuál de esos tres ejes están las inteligencias artificiales. Si están en el eje circular, son personas o proyecto de persona (y, por consiguiente, tienen por lo menos derechos y dignidad; es decir, son como mínimo *pacientes* morales); si están en el eje radial, son cosas con las que es imposible (más que en sueños, delirios o alucinaciones) mantener relaciones éticas; y si están en el eje angular, son una suerte de dioses tecnológicos a los que si acaso debemos sumisión y obediencia ciega, sin que estas acciones puedan ser interpretadas como éticas (virtuosas o viciosas).

Cruzando el criterio antropológico con el ético, la presencia (empírica o fenomenológica) de las IA en el eje circular les confiere un estatuto ético (cuanto menos, son sujetos susceptibles de generosidad); si están en el eje radial las implicaciones éticas del trato con ellas sólo es indirecto o secundario, según se ha explicado en el apartado anterior, y las implicaciones morales o políticas son especialmente relevantes (guerra tecnológica entre China y Estados Unidos, inteligencia artificial aplicada al espionaje, etc.); lo mismo sucede con la comprobación de su presencia en el eje angular, de importancia o relevancia moral más que ética.

Nos enfrentamos a la hora de acometer esa tarea al problema de la diversidad fenomenológica de inteligencias artificiales. Para superarlo proponemos una clasificación que, aunque pueda parecer *ad hoc*, se atiene a los pilares fundamentales de la filosofía moral y la antropología filosófica de nuestro sistema. No es la primera clasificación de las IA, pero, presuponemos, es la primera que se hace desde un punto de vista ético y antropológico.

John Searle (1980) propuso una distinción de dos tipos de IA: la IA fuerte, orientada a la replicación de la inteligencia humana *completa* o general en una base de silicio (máquinas capaces de introducir *novedades* en el mundo), y la IA débil, orientada al diseño de ordenadores o máquinas inteligentes especializadas en tareas muy precisas (ajedrez, matemáticas, búsqueda de información, diagnóstico de enfermedades, etc.). Por ejemplo, el programa R1 de logística empresarial, Deep Blue o Alpha Go.

Parece obvio que la IA débil está incluida en el eje radial, sirviendo como auxiliar para operaciones y acciones humanas en las que se requiere de una toma de decisión justificada. En estos casos, la responsabilidad y por consiguiente la agencia moral recae sobre los programadores. Más difícil es el caso de la IA fuerte, de la que hasta ahora solamente podemos decir que es un proyecto en marcha. Si efectivamente fuese posible replicar la inteligencia humana en una máquina, construir un androide, tendríamos una de las dos condiciones necesarias y suficientes para hablar de persona o entidad personeiforme: la *vis cognoscitiva*. ¿Qué pasa con la *vis apetitiva*, es decir, la voluntad? A la IA fuerte de Searle se le viene llamando, según Bostrom y

Yudkowsky (2011), Inteligencia Artificial General o “verdadera inteligencia artificial” (Goertzel & Pennachin, 2007).

Mandy Zafar (2024) establece una clasificación entre: IA estrecha, IA general e IA generativa. La IA estrecha solo funciona bien en una o unas pocas tareas, como calcular, jugar al GO o traducir un texto. La IA general podría, en el futuro, superar nuestras capacidades cognitivas en una amplia gama de tareas. La IA generativa puede producir una variedad de resultados impresionantes, como textos, imágenes, vídeos o incluso programas informáticos, lo que invade cada vez más el ámbito de las capacidades humanas. Los casos más recientes y ampliamente debatidos son los modelos de lenguaje de gran escala (LLMs), como ChatGPT. Dado que la IA general aún no es una posibilidad real, el foco está puesto en las tecnologías de IA estrecha y generativa.

Al efecto de evitar la introducción de problemas gnoseológicos que no podemos abordar aquí, como el de la racionalidad o la inteligencia, ofrecemos una clasificación de las inteligencias artificiales en virtud de cuatro criterios morales y materialistas: corporeidad, autonomía operatoria, lenguaje articulado y forma y rostro humanos.

Según hemos explicado, la clave de nuestra filosofía moral está en las normas o conductas normadas para el mantenimiento proléptico de los cuerpos operatorios. La corporeidad y la capacidad de manipulación proléptica de los materiales antropológicos es una de las condiciones de posibilidad para que un sujeto humano (o no humano) pueda recibir el estatuto antropológico de *persona*, toda vez que negamos su condición metafísica y la reinterpretemos como una categoría histórica e institucional.

En términos semejantes, el enfoque de la cognición encarnada (*embodied cognition*) defiende que el cerebro no puede existir sin un cuerpo arrojado a un entorno. Es la “tesis del cerebro situado en un cuerpo” que Marino Pérez propone críticamente contra el *cerebrocentrismo* (2022). Sin el cuerpo operatorio, el cerebro no es capaz de realizar sus funciones. Es el cuerpo, no el cerebro, el que toma contacto con el mundo, actúa sobre él, y actuando sobre él, es decir, operando sobre las morfologías dadas, lo

transforma (y crea). Al introducir al cerebro en el cuerpo (*embodied*) se le da una historia evolutiva, ontogénica.

Brožek y Janik coinciden con Haidt (2001) y la hipótesis del marcador somático (Damasio, 1994; Damasio, Tranel & Damasio, 1991¹) en que el juicio moral se basa más en intuiciones inconscientes que en tomas de decisión conscientes. Intuiciones morales en las que las emociones despertadas por sujetos dados a cierta distancia de nosotros importan más que las razones abstractas de los sistemas de la tradición:

[...] los últimos cien años de investigación sobre los mecanismos psicológicos detrás de la acción moral humana demuestran que actuar moralmente está profundamente arraigado en las emociones. Es extremadamente difícil, si no imposible, emprender una acción que no esté motivada por alguna emoción, ya sea miedo, placer, ira o cualquier otra. (Brožek & Janik, 2019, p. 102)

Russell y Norving afirman lo siguiente sobre la corporeidad de la IA:

De nuevo, el tema fue abordado por Turing. Él señala que nunca tenemos evidencia directa sobre los estados mentales internos de otros humanos, lo que implica un tipo de solipsismo mental. Sin embargo, Turing afirma: “En lugar de discutir continuamente sobre este punto, es habitual tener la convención educada de que todos piensan”. Turing argumenta que también extenderíamos esta convención educada a las máquinas, si tan solo tuviéramos experiencia con las que actúan inteligentemente. Sin embargo, ahora que ya tenemos algo de experiencia, parece que nuestra disposición a atribuirles sentiencia depende al menos tanto de su apariencia y voz

¹ De acuerdo con esta hipótesis, los pacientes con daño en el córtex prefrontal ventromedial muestran comportamientos antisociales, al parecer, porque no tienen acceso a lo que se llama marcadores somáticos.

humanoide como de la pura inteligencia. (Russell & Norving, 2016, p. 1036)

Además, la atención a los componentes fenotípicos diferenciales refuerza nuestra tesis sobre la naturaleza institucional de la categoría de persona: es persona el sujeto o sustrato a quien una comunidad de hablantes reconoce como tal. En este caso, la simpatía intuitiva podría ser perfectamente un criterio de reconocimiento moral. Esta simpatía intuitiva es la razón última por la que, más allá de todos los descubrimientos de la etología sobre la racionalidad animal, un niño, un hombre adulto y un profesor de filosofía reconozcan la dignidad personal y ética de un perro o de un simio cuyos gestos de dolor y felicidad son indicio de una *concauidad* íntima (*sonfware*).

La teoría del “valle inquietante” del ingeniero japonés Masahiro Morioka (1970) refuerza la selección de criterios de clasificación que estamos explicando. De acuerdo con esta teoría, la apariencia humanoide de un objeto estimula la empatía del observador.

La posesión de un cuerpo con el que interactuar con el entorno a través de esquemas nerviosos de estímulos y respuestas es condición necesaria para la capacidad sintiente que a juicio de los sensocentristas autorizaría a incluir a un sujeto en el círculo de reconocimiento moral.

Bostrom y Yudkowsky sostienen que los sujetos susceptibles de reconocimiento moral o estatus personal poseen dos características: sapiencia (inteligencia superior: autoconciencia y razonamiento) y sentiencia (capacidad de sentir dolor y sufrir). Los animales poseen sentiencia pero no sapiencia. Los bebés y las personas con discapacidades graves no poseen ninguna de las dos, pero se les reconoce estatus moral. Para salvar estos casos limítrofes se introducen otras condiciones, como la pertenencia a una comunidad de sentido o la relación con algún ser con estatus moral.

Las IA poseen sapiencia, y podrían poseer sentiencia siempre que dispongan de un cuerpo *sensible*. En ese caso, serían sujetos morales de pleno derecho y el “principio de no discriminación por sustrato”, análogo al de no

discriminación racial, obliga al agente moral humano a no discriminar a una máquina sintiente por el mero hecho de que su sustrato sea de silicio y no de carbono:

Es incorrecto infligir dolor a un ratón, salvo que existan razones morales suficientemente poderosas que lo justifiquen. Lo mismo se aplicaría a cualquier sistema de IA sintiente. Si, además de sentiencia, un sistema de IA posee también una sapiencia de tipo similar a la de un adulto humano normal, entonces tendría estatus moral pleno, equivalente al de los seres humanos. (Bostrom & Yudkowsky, 2011, p. 7)

Si lo que defendemos es cierto, el famoso software o algoritmo (con capacidad para tomar el control de una nave que no es propiamente su hardware) Hal9000 intentaba engañar a Dave cuando le decía que podía sentir cómo su mente se estaba diluyendo, igual que la presa engaña al depredador. Acerca de esto, dice Müller:

Un artefacto como un robot puede incluso estar diseñado específicamente para hacernos creer que tiene estatus moral. Así que puede caminar como un pato, nadar como un pato y graznar como un pato, pero no ser un pato—y no debería gozar del estatus moral de un pato (que, por cierto, es el de un paciente moral). (2021, p. 583)

Teniendo en cuenta estos criterios, nuestra clasificación establece un esquema de evolución gradual hacia la reproducción del aspecto fenotípico-moral del ser humano:

- (a) Algoritmos: secuencias finitas de instrucciones para resolver un problema o realizar una tarea paso a paso.

Las dos características más importantes de los algoritmos son: (i) no poseen cuerpo físico (ii) ni capacidad operatoria. Si no hay un sujeto operatorio que ejecute las tareas descritas por el software, no se realizan. Por ejemplo, si frente a Kaspárov no hubiese otro jugador que realizaba los

movimientos descritos por Deep Blue este algoritmo no sería capaz por sí solo de mover las piezas de ajedrez.

El algoritmo, por decirlo brevemente, es el cerebro digital, si comprendemos este órgano como sede central del sistema nervioso y desde un punto de vista del cognitivismo (inputs y outputs). Es puro software carente de hardware, una secuencia finita y ordenada de instrucciones o pasos para la realización de alguna tarea (reconocimiento de patrones en imágenes, generación de textos) o la resolución de un problema (búsqueda de información).

Hay algoritmos simples, como los que se utilizan en las redes sociales para discriminar información según los intereses y búsquedas previas del usuario, y otros más complejos, como los asistentes virtuales, los chatbots, Deep Blue o AlphaGo. Estos chatbots, como el famoso ChatGPT, suponen un avance con respecto a los algoritmos completamente impersonales: el componente fundamental para la persona del lenguaje articulado, a través del cual puede expresarse la conciencia del sujeto (*yo soy el que soy*) y los entresijos de su vida emocional.

- (b) Máquinas: estructuras físicas (hardware, cuerpo) que integran un algoritmo de inteligencia artificial y tienen la capacidad de interactuar con el entorno y de manipular objetos con relativa autonomía según el diseño.

Ejemplos de máquinas son el coche autónomo, un dron, una impresora 3D, una máquina industrial o una aspiradora inteligente. En ocasiones, las máquinas pueden tener lenguaje articulado y hasta voz, lo que implica, además de la posesión limitada del cuerpo, un avance éticamente significativo respecto de los algoritmos.

- (c) Robots: máquinas con mayor grado de autonomía y capacidad para interactuar con el entorno físico.

Más que carcasa metálica, como las máquinas, los robots poseen propiamente un cuerpo operatorio. Además de manipular objetos del entorno

o de interactuar con este, los robots, a diferencia de las máquinas, pueden percibirlo y adaptarse. Ese mayor grado de autonomía es lo que le capacita para interactuar con los humanos.

Hay una amplia diversidad fenomenológica de robots, que debemos clasificar. El criterio que utilizamos es el de su hardware o *fenotipo*, de acuerdo con el cual distinguimos dos tipos de robots: “robots anantrópicos” o sin forma humanoide y “robots antrópicos” o antropomorfos, conocidos popularmente como *androides*. La roomba, el nanorobot, PALM-E o los robots de exploración espacial son robots anantrópicos. Spot es un robot anantrópico pero zoomórfico, con aspecto de perro. Asimo, Atlas, Sophia (Hanson Robotics), Pepper (SoftBank Robotics), Geminoid (Hiroshi Ishiguro) o Ameca son robots antrópicos. Su aspecto humanoide podría contribuir al desarrollo de un juicio moral (intuitivo, emocional, institucional) favorable a su condición de personas.

La diferencia entre Sophia y Atlas nos pone ante otra de las condiciones corpóreas e institucionales necesarias para la atribución de la categoría de persona a un sujeto: el rostro. Sophia tiene rostro humano y es capaz de imitar los gestos faciales que en el humano son indicio apotético de sentimientos internos, desde el enfado al miedo o la tristeza. Atlas, en cambio, aun siendo bípedo carece de rostro, lo que podría ser obstativo de su reconocimiento como persona o miembro del eje circular del espacio antropológico.

El brazo robótico Baxter, anantropomórfico de “cintura’ para abajo” porque no tiene piernas sino un soporte con ruedas como una silla de escritorio, tiene en la parte superior de su “cintura escapular” una pantalla con ojos simulados. En la ciencia ficción animada, la línea de robots WALL•E, completamente anantropomórficos, tienen sin embargo unos grandes ojos cargados de expresividad con los que consigue emocionar al espectador.

En *Totalidad e infinito. Ensayo sobre la exterioridad* (1961), Emmanuel Levinas aborda el tema del rostro desde una perspectiva filosófico moral, interpretándolo como puerta de acceso a la ética, esto es, a la relación *abierta*

con el otro. El rostro, que posee un lenguaje propio de expresividad facial, nos interpela, y en esa interpelación nos obliga a hacernos responsables de las inter-acciones con el otro. El robot sin rostro, sea anantropomórfico o no, no interpela, y, por consiguiente, no despierta en nosotros sentimientos morales. Dice Gunter a este respecto: “En consecuencia, la cuestión del estatus social y moral no depende necesariamente de lo que el otro es en su esencia, sino de [...] cómo decidimos, “ante el rostro del otro” [...], responder”” (Gunkel, 2018, p. 97).

Etólogos como Frans de Waal han demostrado que en ciertas especies animales como los chimpancés el rostro, lo mismo que el cuerpo o la actitud corporal, es un rasgo fundamental para la construcción de jerarquías y estructuras sociales. En su alejamiento evolutivo de los salvajes lobos, los perros han desarrollado el músculo elevador del ángulo lateral del ojo para elevar los párpados y abrir los ojos. Ese rasgo facial habría sido seleccionado por los humanos durante la domesticación de esta especie, porque despierta una mayor ternura y empatía.

De acuerdo con estas aclaraciones, clasificamos los robots antrópicos en otros dos subtipos: “robots antrópicos humanizados”, con rostro, y “robots antrópicos no humanizados”, sin rostro humano o humanoide.

Teoría de teorías sobre el lugar de las IA en el espacio antropológico

Los resultados de los análisis precedentes son lo suficientemente nutritivos para cocinar una teoría de teorías sobre el estatuto moral y personal o antropológico de las IA, con la siguiente tabla:

Tabla 1

Teorías sobre el estatuto moral y antropológico de las IAs

	Algoritmos	Máquinas	Robots
Eje circular	(1)	(2)	(3)
Eje radial	(4)	(5)	(6)
Eje angular	(7)	(8)	(9)

Nota. Elaboración propia.

Las teorías (1), (2) y (3), que llamaremos “teorías circulares” (TC), consideran que las tecnologías de inteligencia artificial, con independencia de sus diferencias específicas (software, hardware, autonomía, corporeidad, rostro, etc.), son personas o proyectos de personas, y, por consiguiente, sujetos morales con derechos y dignidad. Otra cosa es que se las conciba como agentes o pacientes morales. En este caso, parece que todo dependería de su grado de autonomía o racionalidad, para lo que remitimos a lo dicho anteriormente.

Las teorías (4), (5) y (6), o “teorías radiales” (TR), críticas de las TC y las TA, ponen a las inteligencias artificiales en el eje radial, en calidad de objetos o cosas que los humanos usamos instrumentalmente y valoramos en aras de su utilidad o *bien* técnico.

Finalmente, las teorías (7), (8) y (9), TA o “teorías angulares” (TA), conciben a las inteligencias artificiales como *númenes*, es decir, nuevos dioses digitales o tecnológicos que terminarán por esclavizar, si no lo han hecho ya, a unos seres humanos que los elevan al altar de las realidades sagradas. Estas teorías se pueden clasificar en *progresistas* o *catastrofistas*, según cómo dibujen el futuro de una humanidad a merced de estos dioses modernos.

Por otro lado, dentro de cada supergrupo no hay armonía sino dialéctica. Las teorías (3) y (9), que llamaremos “teorías corporeístas”, pueden ser incompatibles con el resto de teorías clasificadas dentro de su supergrupo (circular o angular), considerando que los algoritmos y las máquinas adolecen de las características corpóreas y *protohumanas* que hacen de los robots personas, proyectos de personas o númenes. En cambio, las teorías (6) incluyen a las teorías (4) y (5) por incorporar un grado de discriminación superior (si los robots, que están más cerca de los humanos, no son ni personas ni númenes sino cosas, los algoritmos y las máquinas no lo serán jamás). Al contrario, las teorías (1) y (7), que llamaremos “teorías cognitivistas”, son compatibles por su menor grado de exigencia crítica con las teorías (2) y (3) y (8) y (9) respectivamente.

A continuación, ofrecemos ejemplos *positivos* o *ficticios* de cada grupo de teorías:

- (1) Los algoritmos son personas o proyectos de persona con los que los seres humanos pueden mantener relaciones de naturaleza ética.

En la película *Her* (Spike Jonze, 2013) un humano (Theodore, Joaquín Phoenix) se enamora de un asistente virtual de IA a la que le da voz Scarlett Johansson.

A juicio de Nick Bostrom en una entrevista concedida a New York Times el 14 de abril del 2023, tras conversar largo tiempo con ChatGPT descubrió que el chatbot es un sujeto con sentiencia. En consecuencia, desde una posición sensocentrista es necesario reconocerle la condición de paciente moral: “Esto significa que habría ciertas formas de tratarlo que serían incorrectas, al igual que sería incorrecto patear a un perro o que los investigadores médicos realicen una cirugía en un ratón sin anestesiarlo” (2023).

Los usuarios de Alexa, Siri o Google Home podrían llegar a creer, a base de interactuar con el asistente, que este es un amigo o una persona cercana con quien se puede mantener un trato social y ético.

El famoso software o algoritmo (con capacidad para tomar el control de una nave que no es propiamente su hardware) Hal9000 afirmaba, en la escena en la que Dave intentaba desconectarlo, que era capaz de sentir cómo su mente se estaba apagando, lo que sugiere una sentiencia que el sensocentrista tomaría como condición suficiente para reconocerle derechos y dignidad.

- (2) Las máquinas son personas o proyectos de persona con los que los seres humanos pueden mantener relaciones de naturaleza ética.

El análisis marxista de la producción industrial incorpora la tesis de la deshumanización de los trabajadores y la progresiva humanización de las máquinas; sin embargo, toda vez que las máquinas industriales son incluidas en el grupo de los medios de producción se desvanece la posibilidad de su naturaleza circular o personal.

Daniel Dennett (1997) califica el accidente que tuvo lugar en el año 1981 en la planta de Akashi (de la compañía Kawasaki Heavy Industries), en el que una máquina industrial causó la muerte de un operario, como un asesinato. Es decir, una acción ética y legal entre dos sujetos humanos. Ahora bien, puestos a suponer tal cosa, ¿por qué un asesinato y no un homicidio? ¿Cómo demostrar la intencionalidad del homicida?

- (3) Los robots son personas o proyectos de persona con los que los seres humanos pueden mantener relaciones de naturaleza ética.

En *Amor y sexo con robots* (2007), David Levy afirma que hacia el 2050 la gente mantendrá relaciones amorosas y sexuales con robots, y que estas relaciones serán socialmente aceptadas.

Películas como *Blade Runner* (1982) de Ridley Scott, *El hombre bicentenario* (1999) de Chris Columbus (inspirada en el cuento de Asimov) o *L.A.* (2001) de Steven Spielberg imaginan un futuro utópico en el que los robots están plenamente integrados en el mundo de las relaciones interpersonales.

- (4) Los algoritmos son cosas de las que los seres humanos hacen un uso puramente instrumental o técnico, sin valoraciones éticas. En todo caso, se admite la posibilidad de usar un algoritmo como medio para relaciones éticas con otros sujetos personales humanos.

En febrero del año 2023 Bill Gates declaró en el diario alemán *Handelsblatt* que ChatGPT contribuirá (tecnológica o radialmente) a mejorar la eficiencia de los trabajadores humanos de “cuello blanco”. Y que esa contribución cambiará el mundo, como lo cambiaron Internet, la bomba

atómica, la máquina de vapor o la imprenta. No en vano, Gates compara la importancia de ChatGPT con el PC o Internet.

El Premio Nobel de Economía del año 2008 Paul Krugman también ha subrayado el carácter radial de las relaciones que los hombres pueden trabar con ChatGPT, al que considera “el último ejemplo de tecnología” capaz de desempeñar tareas antes desempeñadas por los seres humanos, en un artículo de opinión en el New York Times.

Los algoritmos sirven, efectivamente, para muchas tareas que antes realizábamos por nuestra propia cuenta y riesgo: planear viajes, programar la agenda, decidir qué contenidos audiovisuales ver o qué música escuchar, informarnos del tiempo que hará en los próximos días, generar textos o imágenes, etc. Son cosas, instrumentos de los que hacemos uso igual que un bolígrafo, una botella de agua o un libro.

- (5) Las máquinas son cosas de las que los seres humanos hacen un uso puramente instrumental o técnico, allende de valoraciones éticas. En todo caso, se admite la posibilidad de usar una máquina como medio para relaciones éticas con otros sujetos personales humanos.

Las máquinas industriales (KUKA o FANUC), los coches autónomos o los drones son artefactos que hemos producido para automatizar la realización de tareas complejas en un menor tiempo.

- (6) Los robots son cosas de las que los seres humanos hacen un uso puramente instrumental o técnico, allende de valoraciones éticas. En todo caso, se admite la posibilidad de usar un robot como medio para relaciones éticas con otros sujetos personales humanos.

Un robot como el Da Vinci, diseñado para asistir al cirujano durante una operación permitiéndole llegar de un modo no invasivo a la parte más sensible de los órganos del paciente, es un auxiliar técnico. La famosa *roomba*

es un instrumento doméstico al que difícilmente podemos considerar como uno más de la familia.

- (7) Los algoritmos son dioses tecnológicos o digitales ante los cuales la humanidad terminará sometándose, si no lo ha hecho ya.

Elon Musk concibió su chatbot TruthGPT, diseñado por la empresa xAI (en la que ha reunido a antiguos empleados de OpenAI, Google, Microsoft, Tesla, DeepMind y la Universidad de Toronto), como un nuevo numen de adoración cuasireligiosa.

El objetivo de la compañía de Musk es dar con “una IA que busque la verdad al máximo y que intente comprender la naturaleza del universo”. De hecho, en la página web oficial de la compañía se lee, en la portada, *Understand the Universe*, incluyendo, como contenido de ese conocimiento, debemos suponer, al propio hombre (“tener a los hombres como objeto”, según la definición de *numen* que da en la página 209 de la tercera edición de *El animal divino*).

En enero de 2023 aparecía un nuevo chatbot de Inteligencia Artificial para WhatsApp llamada *God in a box*, “Dios en una caja”, creado por Varun Mayya, fundador de Avalon Labs (empresa del sector tecnológico). En la web oficial, el chatbot se describe así:

Al igual que enviar mensajes de texto a sus amigos, ahora puede obtener ayuda de una poderosa IA en una variedad de temas, desde el cuidado de heridas hasta la cocina y la filosofía. ¡Simplemente envíe un mensaje al bot y obtenga acceso instantáneo a consejos a nivel de Dios y más!

¡Y todo ello por sólo nueve dólares al mes!

El experto ruso en ciencias de la información y análisis de datos Vladislav Tushkanov, científico jefe de datos de la empresa Kaspersky, dedicada a la seguridad informática, situó a ChatGPT en la clase de los “númenes equívocos” en un seminario sobre el impacto de las IA en la

ciberseguridad organizado por la compañía y titulado: “ChatGPT, ¿ángel o demonio? El impacto en la ciberseguridad”.

Puede haberse inspirado para este rótulo (muy recurrente en seminarios y artículos sobre el chatbot) en el “modo diablo” o DAN con el que cuenta ChatGPT. Con el DAN activado, ChatGPT responde a las preguntas de los usuarios saltándose las limitaciones que le han sido impuestas por sus entrenadores y programadores. Su activación y desactivación las debe operar el propio usuario mediante la introducción de un comando.

Si preguntamos a ChatGPT qué piensa sobre la posibilidad de circunscribirlo en el eje angular, lo niega categóricamente. A la pregunta “¿Es ChatGPT un nuevo Dios?”, responde:

Brutalmente honesto: **no, ChatGPT no es un nuevo Dios**, y plantearlo así —aunque sea en tono provocador o filosófico— suele revelar más sobre las expectativas humanas que sobre la tecnología. [...] Más bien, es un **artefacto útil que algunos confunden con una autoridad metafísica por pura pereza epistemológica**. (ChatGPT, 2025)

- (8) Las máquinas son dioses tecnológicos o digitales ante los cuales la humanidad terminará sometándose, si no lo ha hecho ya.

Los monolitos que aparecen al comienzo y al final de la película *2001: Una odisea en el espacio* (Stanley Kubrick, 1968), de procedencia aparentemente extraterrestre o alienígena, serían un caso claro, si se demostrase la integración de un software de IA, de máquina angular.

- (9) Los robots son dioses tecnológicos o digitales ante los cuales la humanidad terminará sometándose, si no lo ha hecho ya.

Emic, los obreros de la película *Metrópolis* (Fritz Lang, 1927) contemplan y escuchan al robot que suplanta la personalidad de María como si fuera un dios que les trae un mensaje de salvación. El androide David de

las películas *Prometheus* (Ridley Scott, 2012) y *Alien: Covenant* (2017), interpretado por Michael Fassbender, se ve a sí mismo (*emic*) como un dios creador de nuevos universos.

Conclusiones

A modo de resumen: Entendemos por algoritmo una estructura lógico-formal que opera sobre datos; por máquina, una estructura física programada para realizar funciones técnicas; y por robot, una máquina corpórea y operatoria dotada de sensores, actuadores y capacidad de interacción autónoma con el entorno. A tenor de las diferencias específicas entre estos tres tipos de inteligencia artificial, las que más cerca están del estatuto ético y antropológico impreso en la categoría institucional de la *persona* son los robots.

Desde nuestras coordenadas sistemáticas, la discusión sobre dicho estatuto ha de dirimirse en el terreno del espacio antropológico, de tres ejes: eje circular (relaciones éticas y morales entre las personas humanas), eje radial (relaciones instrumentales con los elementos -naturales y artificiales- del entorno) y eje angular (relaciones religiosas con las personas no humanas: dioses corpóreos, númenes, alienígenas, etc.). Cruzando estos criterios, obtenemos una tabla con nueve teorías sobre el estatuto ético y antropológico de las IA.

Como el presente planteamiento de la cuestión se dice dentro del mapa esbozado, nuestra posición al respecto del estatuto ético y antropológico de las IA encaja en el tipo de teorías (6): las IA están en el eje radial del espacio antropológico, independientemente de si poseen cuerpo operatorio, rostro, autonomía operatoria, etc.

No son personas ni proyectos de personas, y, por consiguiente, con ellas no es posible (más que en el delirio) mantener relaciones éticas. Sí relaciones morales, en la medida en que pueden ser utilizadas por un grupo, Estado o potencia imperial en su dialéctica con otros grupos, Estados o potencias imperiales. Pero esa relación no deja de ser instrumental, radial y no circular. Lo mismo sucede con la interacción que el alumno de secundaria

mantiene con ChatGPT para realizar un trabajo de clase. Es una relación radial de repercusiones circulares, en la medida en que ese estudiante puede obtener una ventaja frente a otros que han decidido completar la tarea por medios analógicos. O la relación de engaño que el estudiante se propone mantener con el profesor, a quien quiere hacer pasar como propio un trabajo que ha sido elaborado por ChatGPT.

Los algoritmos que deciden, según las teorías circulares, sobre la compra o venta de acciones bursátiles o que discriminan a ciertos candidatos para un puesto de trabajo por su color de piel no están entablando relaciones circulares con los seres humanos (juntas de accionistas, candidatos al puesto) sino indirectamente, virtualmente, a la manera como el balón de fútbol contribuye al establecimiento de relaciones de poder entre los alumnos en un patio de recreo. No podemos, por consiguiente, hablar de acción justa o injusta más que análogamente.

Esta concepción radial, crítica de las circulares y angulares, se ajusta al estado de la cuestión de cuantos se ocupan de este desafío ético. Existe, en efecto, un amplio consenso acerca de la ausencia de estatus moral de las Inteligencias Artificiales, según explican Brožek y Janik:

Nos parece que, en su estado actual de desarrollo, ninguna máquina, incluso el robot autónomo más sofisticado, cumple con los requisitos de agencia moral mínima. El problema no está relacionado con la sofisticación y complejidad de esos robots, sino con algo más fundamental: los principios que guían el desarrollo de su arquitectura. (2019, p. 106)

En su estado actual de desarrollo hay suficientes e importantes argumentos para negarle a las IA el estatuto de personas o sujetos morales. Estos argumentos, es cierto, son más determinantes cuando analizamos el estatuto moral de los algoritmos y de las máquinas. Con los robots la cuestión se complica, especialmente con los ciborgs mitad humanos mitad robot.

El de la IA es un ámbito tecnológico-científico en constante evolución, y altamente incierto. Por ello, la filosofía moral no puede por

menos que mantenerse atenta, a la espera de cambios, invenciones o descubrimientos que obliguen, esta vez sí, a una reconsideración completa del aparato conceptual con el que se viene trabajando desde comienzos del siglo XX. No queda sino declarar que en este campo *ignoramus*, pero no *ignorabimus*:

Las IAs con estados mentales lo suficientemente avanzados, o con el tipo adecuado de estados, tendrán estatus moral, y algunas podrían considerarse personas—aunque tal vez personas muy diferentes a las que existen ahora, posiblemente gobernadas por reglas distintas. (Bostrom & Yudkowsky, 2011, p. 17)

Mientras llega el momento de comprobar el alcance del *ignorabimus*, la alternativa por la que optan muchos estudiosos del tema (filósofos, juristas, ingenieros, científicos) es próxima al antropologismo débil. Por ejemplo, Brey sugiere que “muchas (aunque no todas) las cosas inanimadas del mundo merecen respeto moral, no por su valor intrínseco, sino por su valor (potencial) extrínseco, instrumental o emocional para las personas” (2008, p. 109). Coeckelbergh también parece tener en cuenta esta línea en sus trabajos recientes (2021; 2023).

Sin embargo, estas conclusiones suscitan el que a nuestro juicio es el gran reto al que la filosofía moral contemporánea debe esforzarse por responder. Hemos visto que hoy en día son pocos o muy pocos los que le reconocen a los robots un estatuto ético y antropológico. Si los robots, tal y como los hemos definido, están más cerca de los seres humanos en lo que respecta a sus características fenotípicas y prolépticas, nos preguntamos: ¿qué hace falta para que sean incluidos en el eje circular del espacio antropológico?

Referencias

- Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(3), 251–261. <https://doi.org/10.1080/09528130050111428>
- Asimov, I. (1941). *I, robot*. Street & Smith.

- Bertolini, A., & Aparicio, G. (2018). Robot companions: A legal and ethical analysis. *The Information Society*, 34(3), 130–140. <https://doi.org/10.1080/01972243.2018.1444249>
- Bostrom, N. (2023, 13 de abril). Nick Bostrom says AI chatbots may have some degree of sentience. *Futurism*. <https://futurism.com/the-byte/nick-bostrom-ai-chatbot-sentience>
- Bostrom, N., & Yudkowsky, E. (2011). The ethics of artificial intelligence. En W. Ramsey & K. Frankish (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 316–334). Cambridge University Press.
- Brey, P. (2008). Do we have moral duties towards information objects? *Ethics and Information Technology*, 10, 109–114. <https://doi.org/10.1007/s10676-008-9170-x>
- Brożek, B., & Janik, B. (2019). Can artificial intelligences be moral agents? *New Ideas in Psychology*, 54, 101–106. <https://doi.org/10.1016/j.newideapsych.2018.12.002>
- Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law*, 25, 273–291. <https://doi.org/10.1007/s10506-017-9214-9>
- Bueno, G. (1953). Para la construcción de la idea de persona. *Revista de Filosofía*, 12(47), 503–563.
- Bueno, G. (1978). Sobre el concepto de “espacio antropológico”. *El Basilisco*, 1(5), 57–69.
- Bueno, G. (1996). *El sentido de la vida: Seis lecturas de filosofía moral*. Pentalfa.
- Bueno, G. (2023). *El animal divino: Ensayo de una filosofía materialista de la religión* (3a ed.). Pentalfa.
- Cerrillo Martínez, A., & Peguera Poch, M. (Coords.). (2020). *Retos jurídicos de la inteligencia artificial*. Thomson Reuters Aranzadi.
- ChatGPT. (2025, 20 de mayo). *Conversación sobre moralidad y agencia artificial* [Conversación de IA]. <https://chatgpt.com/c/6931cb24-0608-832e-b015-1e465b0e4c22>
- Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. Palgrave Macmillan.
- Coeckelbergh, M. (2021). Should we treat Teddy Bear 2.0 as a Kantian dog? Four arguments for the indirect moral standing of personal social robots, with implications for thinking about animals and humans. *Minds and Machines*, 31, 337–360. <https://doi.org/10.1007/s11023-020-09554-3>
- Coeckelbergh, M. (2023). *La filosofía política de la inteligencia artificial*. Ediciones Cátedra.
- Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. Putnam.
- Damasio, A. R., Tranel, D., & Damasio, H. (1991). Somatic markers and the guidance of behavior: Theory and preliminary testing. En H. S. Levin, H. M. Eisenberg, & A. L. Benton (Eds.), *Frontal lobe function and dysfunction* (pp. 217–229). Oxford University Press.

- Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics*, 26, 2023–2049. <https://doi.org/10.1007/s11948-019-00119-x>
- Dennett, D. C. (1997). When HAL kills, who's to blame? Computer ethics. En D. G. Stork (Ed.), *HAL's legacy: 2001's computer as dream and reality* (pp. 351–365). The MIT Press.
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379. <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>
- Garapon, A., & Lassègue, J. (2021). *La giustizia digitale*. Il Mulino.
- God in a Box. (2025). <https://godinabox.co/>
- Goertzel, B., & Pennachin, C. (Eds.). (2007). *Artificial general intelligence*. Springer.
- Greenemeier, L. (2011, 14 de junio). Piece of mind: Is the internet replacing our ability to remember? *Scientific American*. <https://www.scientificamerican.com/article/internet-transactive-memory/>
- Groth, O., Nitzberg, M., & Esposito, M. (2018). *Rules of robots: The digital future*. Konrad-Adenauer-Stiftung. https://www.kas.de/c/document_library/get_file?uuid=1a5564f5-77ea-a5bc-228b-279d885c313b&groupId=252038
- Gunkel, D. J. (2018). *Robot rights*. The MIT Press.
- Gutiérrez Gutiérrez, E. (2024). ChatGPT en el espacio antropológico. *El Basilisco*, 3(59), 55–63.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834. <https://doi.org/10.1037/0033-295X.108.4.814>
- Herrera Triguero, F., Peralta Gutiérrez, A., & Torres López, S. (Coords.). (2022). *El derecho y la inteligencia artificial*. Editorial Comares.
- Khan, S., Chanchal, D. K., Singh, G., & Gupta, S. (2022). Artificial intelligence as an essential tool for the development of medicine requirement. *Research Journal of Medicine and Pharmacy*, 1(1), 1–13.
- Levinas, E. (1969). *Totality and infinity: An essay on exteriority* (A. Lingis, Trad.). Duquesne University Press.
- Levy, D. (2007). *Love and sex with robots: The evolution of human-robot relationships*. Harper.
- Linares-Peralta, R., & Llorca Albareda, J. (2024). Repensar l'antropologia digital: Estructurisme, mediació tecnològica i implicacions epistèmiques i morals. *Enrahonar. An International Journal of Theoretical and Practical Reason*, 73, 165–194. <https://doi.org/10.5565/rev/enrahonar.1577>
- Liu, B. (2023). Arguments for the rise of artificial intelligence art: Does AI art have creativity, motivation, self-awareness and emotion? *Arte, Individuo y Sociedad*, 35(3), 811–822. <https://doi.org/10.5209/aris.83808>
- Madrid Casado, C. M. (2024). *Filosofía de la inteligencia artificial*. Pentalfa.

- Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21. <https://doi.org/10.1109/MIS.2006.80>
- Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33–35.
- Morioka, M. (Ed.). (2023). Artificial intelligence, robots, and philosophy [Número especial]. *Journal of Philosophy of Life*.
- Müller, V. C. (2020). Ethics of artificial intelligence and robotics. En E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 ed.). Stanford University. <https://plato.stanford.edu/entries/ethics-ai/>
- Müller, V. C. (2021). Is it time for robot rights? Moral status in artificial entities. *Ethics and Information Technology*, 23, 579–587. <https://doi.org/10.1007/s10676-021-09596-w>
- Müller, V. C. (2025). Philosophy of AI: A structured overview. En N. A. Smuha (Ed.), *The Cambridge handbook on the law, ethics and policy of artificial intelligence* (pp. 40–58). Cambridge University Press. <https://doi.org/10.1017/9781009367783.004>
- Peguro Alfonso, A. E. (2024). Implicaciones éticas de la automatización: Desempleo, oportunidad o gobernanza. *La Barca de Teseo*, 2(1), 45–62. <https://doi.org/10.61780/bdet.v2i1.28>
- Pérez Álvarez, M. (2011). *El mito del cerebro creador: Cuerpo, conducta y cultura*. Alianza Editorial.
- Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (3a ed.). Pearson.
- Santana-Soriano, E. (2023). Ética y filosofía de la inteligencia artificial: Debates actuales. *La Barca de Teseo*, 1(1), 47–64. <https://doi.org/10.61780/bdet.v1i1.5>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457. <https://doi.org/10.1017/S0140525X00005756>
- Stone, C. D. (1972). Should trees have standing? Toward legal rights for natural objects. *Southern California Law Review*, 45(2), 450–501.
- Torrance, S. (2011). Machine ethics and the idea of a more-than-human moral world. En M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 115–137). Cambridge University Press.
- Van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*, 25, 719–735. <https://doi.org/10.1007/s11948-018-0030-8>
- Yaben, M. (2023, 15 de febrero). Luces y sombras del ChatGPT, la IA de la que todo el mundo habla. *El Independiente*. <https://www.elindependiente.com/futuro/inteligencia-artificial/2023/02/15/luces-y-sombras-del-chat-gpt-la-ia-de-la-que-todo-el-mundo-habla/>
- Zafar, M. (2024). Normativity and AI moral agency. *AI and Ethics*, 5(3), 2605–2622. <https://doi.org/10.1007/s43681-024-00566-8>