Ética y filosofía de la inteligencia artificial: debates actuales

Ethics and Philosophy of Artificial Intelligence: Current Debates

Santana-Soriano, Edwin
Universidad del País Vasco, España
Universidad Autónoma de Santo Domingo, República Dominicana
esantana004@ikasle.ehu.eus
ORCID: https://orcid.org/0000-0002-4314-6531

Recibido: 2023/04/13 **Aceptado**: 2023/10/03 **Publicado**: 2023/11/22

RESUMEN

Este documento ofrece un análisis crítico y sistemático de la literatura pertinente en torno a las discusiones éticas de los recientes avances en inteligencia artificial (IA). Con el objetivo de profundizar en las implicaciones éticas de la IA, se pone especial énfasis en sus aplicaciones en los ámbitos de la salud y la política, destacados por su importancia intrínseca en la estructuración de la vida cotidiana y la cohesión social. La discusión se centra en la influencia de la IA en aspectos esenciales como la toma de decisiones y la privacidad en salud; mientras que, en el contexto político, se subraya su potencial impacto en la opinión pública, la propagación de desinformación y la vulnerabilidad de las instituciones ante ciberataques. Se concluye que, en tanto la integración de la IA en diversos ámbitos ha suscitado preocupaciones éticas legítimas y ha presentado riesgos políticos concretos, se hace necesario fomentar el diálogo interdisciplinario para anticipar y comprender el horizonte tecnológico emergente y establecer de manera clara los principios que guiarán su concepción, producción, desarrollo, implementación, distribución y uso. Adicionalmente, se argumenta que, si bien la IA es capaz de emular ciertos procesos cognitivos humanos, su actual incapacidad para experimentar emociones o autoconciencia la distingue fundamentalmente de la inteligencia humana, cuestionando cualquier comparación directa entre ambas.

PALABRAS CLAVE

Filosofía de la inteligencia artificial, ética, filosofía de la tecnología, salud, política

ABSTRACT

This paper provides a critical and systematic analysis of the relevant literature concerning the ethical discussions of recent advancements in artificial intelligence (AI). Aiming to delve deeper into the ethical implications of AI, particular emphasis is placed on its applications in the fields of health and politics, both underscored for their intrinsic significance in shaping daily life and social cohesion. The discussion focuses on the influence of AI on key aspects such as decision-making and privacy within healthcare. In the political context, attention is drawn to its potential impact on public opinion, the spread of misinformation, and the vulnerability of institutions to cyberattacks. It is concluded that as the integration of AI across various domains has raised legitimate ethical concerns and posed tangible political risks, there is a pressing need to promote interdisciplinary dialogue to anticipate and understand the forthcoming technological horizon and to clearly establish the guiding principles for its conception, production, development, implementation, distribution, and use. Furthermore, it is contended that while AI can emulate certain human cognitive processes, its current inability to experience emotions or self-awareness fundamentally distinguishes it from human intelligence, challenging any direct comparison between the two.

KEYWORDS

Philosophy of artificial intelligence, ethics, philosophy of technology, health, politics

1. INTRODUCCIÓN

La noción de una esencia humana inmutable ha sostenido las bases éticas de Occidente hasta fechas recientes, y frecuentemente, esta noción ha desestimado lo artificial. Sin embargo, en la actualidad, nos encontramos frente a un viraje en el panorama: la ciencia y la tecnología se consolidan como poderosas palancas del avance humano, tanto en el ámbito ontológico como en el epistemológico y el ético, afectando, no sólo su ser, sino su saber, sus relaciones y la convivencia entre sujetos y con su entorno. En el presente ensayo, si bien se reconoce y valora el papel crucial de la tecnología y la ciencia, se subraya con rigor crítico, la necesidad de escrutar y discernir sobre sus aplicaciones, poniendo especial énfasis en las más recientes y asombrosas conquistas tecnológicas, especialmente en términos de la denominada inteligencia artificial y su relación con la ética, más específicamente aún en lo que respecta a la gestión de la salud y la política, como áreas de especial interés para una relación armoniosa entre los individuos que componen la sociedad y de estos con el ambiente que los alberga.

La inteligencia artificial es un campo de estudio que surge en la década de 1950 en el seno de la informática, la ciencia del tratamiento de la información y la comunicación para fines aplicados o técnicos. No obstante, por la amplitud aspiracional de aquel campo de estudio, que no es menos que lograr máquinas que sean capaces de emular las actividades humanas que se consideran “inteligentes”, el de la Inteligencia Artificial (IA), no es campo unidisciplinar, sino que se nutre activamente de todos los saberes que de alguna manera puedan aportarle a la consecución de tan ambiciosa meta, entre lo que se encuentran, de una manera muy fundamental, la matemática, la lógica y la neurociencia (Merejo, 2023a; 2023b).

El presente es un análisis, cuyo objetivo principal es aportar a la comprensión de las implicaciones éticas y filosóficas de los avances en inteligencia artificial, especialmente en las ciencias de la salud y la política, retoma y pone de relieve los desafíos y oportunidades que la IA presenta para el bienestar y la convivencia de los individuos y la sociedad en su conjunto. En esta búsqueda, se hace un acercamiento a los desafíos inherentes a este campo de investigación, aunados a la discusión multidisciplinar sobre la IA y su relación con la inteligencia humana y los fundamentos filosóficos detrás de la IA. Y se analizan, además, con el apoyo de la perspectiva y la crítica de otros autores e investigadores de diversas áreas, las aplicaciones prácticas de la IA en medicina y política, y las consideraciones éticas en estas áreas.

Se trata, pues, de un análisis crítico de la literatura relevante en torno a esas cuestiones filosóficas que guardan relación con los avances recientes en inteligencia artificial, que se llevó a cabo luego de una búsqueda en las bases de datos académicas más relevantes y en revistas especializadas, priorizando textos

de la última década y aquellos escritos por autores de renombre en cada campo. Una vez identificados los textos relevantes para este diálogo, se procedió a extraer sus argumentos y posturas centrales. Esta información, que posteriormente fue comparada y analizada críticamente identificando patrones y discrepancias en las reflexiones presentadas, permitió la elaboración de una síntesis que, a su vez, permite inferir las conclusiones sobre la problemática que aquí se presenta.

2. LA FILOSOFÍA DETRÁS DE LA INTELIGENCIA ARTIFICIAL

Desde el punto de vista teórico, la mayor dificultad para que sea alcanzado el objetivo en el seno de la inteligencia artificial en tanto campo de investigación, a nuestro modo de ver, viene dada por dos factores principales, a saber: la complejidad de la definición de inteligencia y la imposibilidad de acuerdo en torno a las orientaciones éticas y axiológicas por parte de los desarrolladores de tecnologías y la comunidad científica y académica. Otros factores importantes a tener en cuenta son los ingentes recursos de diversos tipos que se necesitan para el desarrollo de estas tecnologías -factor de alta incidencia en el debate, puesto que podría imponerse por encima de los primeros dos mencionados-, y los intereses políticos que pudieran rodear estos desarrollos. Empero, en el presente trabajo nos vamos a enfocar en los primeros dos: la complejidad de la definición de inteligencia y la imposibilidad de acuerdo en torno a las orientaciones éticas que deben guiar el desarrollo de tecnologías de inteligencia artificial.

El debate sobre el primer factor se trabaja dentro de los límites de la filosofía de la IA, que se encarga de responder a la pregunta de qué es la inteligencia no biológica y qué la hace posible, constituyéndose en una base filosófica y metodológica para el estudio de esa "inteligencia no biológica" (Bezlepkin & Zaykova, 2021). Sin embargo, este mismo debate se remonta a discusiones más fundamentales, como el problema de la mente, su ontología y su relación con el cuerpo, y la definición de la conciencia.

Históricamente, ese problema relacionado con las dicotomías mente-cerebro, cuerpo-alma, yo-cerebro, se ha encarado y se puede encarar aún, bajo tres marcos actitudinales: descartar la relevancia del problema por considerarlo un pseudoproblema, aceptar que se trata de un problema pero que no tiene solución o aceptar que es un auténtico problema con posibles soluciones. Las primeras dos actitudes no enfrentan la situación, aunque en el segundo caso, por lo menos, se le analiza antes de cortar toda esperanza. Pero la tercera actitud, en tanto entiende que se está frente a un problema con posibles soluciones, permite, y ha permitido, desarrollar el debate, dejando abierta la posibilidad de hallar respuestas a las cuestiones que allí se presentan.

Un ejemplo de la primera actitud lo constituyen las posturas conductistas y de los positivistas lógicos -hablando en términos generales, puesto que hay excepciones-, pues estos partían del supuesto de que lo único estudiable científicamente, en esta supuesta relación (mente-cerebro) eran los actos conductuales. De la segunda actitud puede mencionarse como ejemplos a David Hume y, posteriormente, a Herbert Spencer quienes, entre otros, se fundamentaban en un *ignoramus et ingnorabimus:* no sabemos cómo es que surgen esos fenómenos mentales a partir de actividades propias del cuerpo y, en especial del cerebro, ni estamos en condiciones de saberlo en el futuro (Cfr. Bunge, 1985).

La tercera actitud, que por obvias razones puede afirmarse que posee una potencialidad heurística superior a las otras, y que genera debates intensos e interesantes, puede a su vez ser objeto de división en dos partidos: aquellos que defienden que el elemento responsable de las actividades que se denominan "mentales" (pensar, desear, percibir, etc.) es la mente, o el alma o el espíritu; y aquellos que defienden la idea de que esas actividades "mentales" las realiza un complejísimo órgano material ubicado en el cuerpo, llamado cerebro.

Desde ese debate puede vislumbrarse la base filosófica detrás del impulso de la IA como campo de estudio. Pues, aunque los partidarios del primer grupo asumen que las actividades denominadas mentales -y que definen a los seres humanos como seres inteligentes- son realizadas por una entidad inmaterial que produce los sentimientos, recuerdos, e ideas, los filósofos, científicos e ingenieros que defienden la otra postura, parten del hecho de que eso a lo que se le denomina "mente" no es una entidad independiente del cerebro, sino que se trata de una colección de actividades y funciones cerebrales precisamente, que permiten al sujeto en conjunto recordar, imaginar, percibir y emocionarse, entre otras cosas. Y son los defensores de esta vía de solución al problema mente-cuerpo los que están en disposición de, al menos, intentar emular las características que hacen que un ser humano pueda ser considerado inteligente, de manera artificial. Es decir, que son estos filósofos, científicos e ingenieros que entienden que la mente puede ser explicada a partir de funciones del cerebro, los que opinan y defienden que es posible, en principio, construir una inteligencia no biológica.

Tal y como lo definen Bezlepkin y Zaykova (2021), esa inteligencia no biológica habría de contener ciertas características distintivas, como capacidad de procesamiento de información, capacidad de aprendizaje y capacidad de adaptación al entorno. Y la idea motriz de la IA como campo de estudio radica precisamente en que se entiende que esto se puede lograr a través de la creación de sistemas informáticos, como es el caso de las denominadas redes neuronales artificiales, que son sistemas de procesamiento de información que se inspiran en el funcionamiento del cerebro humano y están diseñados para aprender y adaptarse a partir de ejemplos y datos. Se trata de redes compuestas por nodos o neuronas artificiales

que están conectados entre sí mediante enlaces o conexiones, en las que cada neurona artificial recibe una o varias entradas, las procesa y produce una salida que se transmite a otras neuronas. En efecto, esas redes neuronales artificiales se utilizan en la actualidad en una amplia variedad de aplicaciones, como el reconocimiento de patrones, la clasificación de datos, la predicción y la toma de decisiones.

Este desarrollo se fundamenta, tácita o explícitamente, en una definición de inteligencia que surge de la premisa de que es factible construir una entidad que manifieste todas las características inherentes al ser humano, incluidas aquellas más controvertidas, para ser clasificada como inteligente. No obstante, incluso las conceptualizaciones más recientes sobre la inteligencia pueden ser clasificadas en, al menos, tres enfoques distintos, basados en el criterio predominante, a saber: la capacidad cognitiva, la función global y la resolución de problemas junto con la creatividad.

Un ejemplo del enfoque que enfatiza la capacidad cognitiva al definir la inteligencia es la propuesta articulada por Alonso-Fernández (1996). Este autor conceptualiza la inteligencia como "un acto intrapsíquico de comprensión racional", subrayando elementos clave tales como: a) la aptitud para enfrentar desafíos inéditos; b) la utilización constructiva del pensamiento; c) la facultad de síntesis; y d) la distinción crítica entre elementos esenciales y periféricos. Esta perspectiva, como se observa, pone de relieve las habilidades mentales y el proceso de razonamiento, enfocándose en la habilidad del sujeto para procesar y asimilar información de manera eficaz.

En cuanto al enfoque de función global, se pone énfasis en la manifestación de la inteligencia en la conducta integral del ser humano. Un ejemplo de este es el de Adams et al. (1997), quienes conceptualizan la inteligencia como un "agregado o capacidad global de un individuo de actuar con un propósito, de pensar racionalmente y de manejar eficientemente su ambiente externo". Una visión que sugiere que la inteligencia trasciende habilidades puntuales, planteándola como una competencia bastante general que determina cómo el individuo se relaciona con su entorno.

Finalmente, el tercer enfoque, centrado en la resolución de problemas y creación, pone de manifiesto la aplicación práctica de la inteligencia. Un ejemplo paradigmático es el de Howard Gardner (1997), quien define la inteligencia como "la capacidad de resolver problemas o crear productos que sean valiosos en uno o más ambientes culturales", subrayando la importancia de la aplicabilidad y la pertinencia cultural de la inteligencia, y sugiriendo que la verdadera esencia de la inteligencia no radica únicamente en la capacidad cognitiva, sino en cómo esta se materializa en escenarios concretos.

No obstante, el concepto de inteligencia artificial parece apoyarse en definiciones más operativas de inteligencia, como la que puede inferirse -guardando las distancias- a partir de René Descartes y su

Tratado del hombre, cuando equipara el funcionamiento del cuerpo al de una máquina. Aunque Descartes (1664) argumentaba que los animales eran como máquinas, movidos por mecanismos simples, mientras que los humanos tenían una mente racional capaz de realizar cálculos complejos, la idea posterior de que un software puede hacer funcionar un hardware bien podría hallar en Descartes a un precursor, por rebuscada que nos parezca la idea.

Otra definición operativa puede hallarse en Leibniz quien, aunque no afirmó explícitamente que la mente humana es una calculadora, sí creía que los procesos mentales y lógicos podían ser representados y comprendidos en términos de cálculos y, además, sus ideas y trabajos en lógica, matemáticas y filosofía sentaron las bases para desarrollos posteriores en la teoría de la computación y la lógica formal (Freidman, 2023; Di Liscia y Sylla, 2022).

Ya en el siglo XX, con el auge de la computación, la idea de inteligencia como capacidad de cálculo ganó más terreno. Alan Turing, en su famoso artículo *Computing Machinery and Intelligence* (1950), no sólo propuso la idea de una máquina que podría imitar la inteligencia humana, sino que planteó preguntas fundamentales sobre qué significa pensar. Turing (1950) introdujo el ahora famoso "Test de Turing" como una medida para determinar si una máquina podía ser considerada "inteligente" y, sobre esa idea, construyeron McCarthy y Hayes (1981), cuyos trabajos dan por sentado que una máquina puede ser inteligente en el mismo sentido que lo es un ser humano o un animal y se enfocan en desarrollar una representación formal del mundo y del razonamiento que permita a una máquina actuar inteligentemente en ese mundo.

De modo sintético, podría explicitarse la definición operativa tácita de inteligencia de trasfondo en la inteligencia artificial como sigue: la inteligencia es la capacidad de planificar, aprender razonar, percibir e interactuar con el entorno; y esto puede hallar justificación en trabajos más recientes que, teniendo como base los anteriores citados, permiten construir esa definición.

Por ejemplo, Ghallab et al. (2004) aseguran que las máquinas planifican porque utilizan algoritmos de planificación que les permiten decidir la secuencia de acciones necesarias para alcanzar un estado objetivo. Esto basados en la definición de planificar como capacidad de establecer y seguir un conjunto de pasos para alcanzar un objetivo específico.

En el caso de aprender, en estos términos operativos ha sido definido como la capacidad de adquirir conocimiento o habilidades a través de la experiencia, el estudio o la enseñanza. Partiendo de esa definición, Russell y Norvig (2010), en *Artificial Intelligence: A Modern Approach,* afirman que, en efecto, las máquinas aprenden a través de algoritmos de aprendizaje automático que permiten que se les

proporcione una gran cantidad de datos y se les programa para, o enseña a, reconocer patrones y hacer predicciones basadas en esos datos.

En esa línea, Marr (1982) ya había asegurado que las máquinas en efecto podían percibir a través de sensores y algoritmos de procesamiento de señales que les permiten interpretar datos del mundo real, como imágenes y sonidos, partiendo de la definición de percibir como la capacidad de reconocer y responder a estímulos del entorno.

Del mismo modo, la definición de razonar que se usa es aquella que la asume sólo como la capacidad de pensar lógicamente y tomar decisiones basadas en la información disponible. Esta definición le permite a Haugeland (1985), para poner un ejemplo, afirmar que las máquinas realizan esta actividad cuando, utilizando lógica formal y algoritmos de inferencia, pueden llegar a conclusiones basadas en premisas.

Así mismo, partiendo de la definición de interactuar como capacidad de un ente de comunicarse y actuar en relación con otros entes o sistemas, Winograd y Flores (1986) afirman que las máquinas interactúan cuando, utilizando interfaces de usuario y protocolos de comunicación, pueden recibir entradas y enviar salidas a otros sistemas o usuarios.

Como se ha visto hasta aquí, la concepción de que una máquina puede ser programada para emular procesos cognitivos humanos ha promovido la creación de sistemas que se reputan como capaces de aprender, razonar, planificar, percibir y, en algunos casos hasta se les atribuye la capacidad de interactuar con el entorno, utilizando conceptos que otrora fueran para referencia casi exclusiva de los seres humanos.

Empero, la discusión se mantiene en los ambientes académicos, puesto que se pone en duda si esa definición de inteligencia sea suficiente como para comprender la inteligencia humana, toda vez que, en muchas ocasiones, se intenta equiparar la inteligencia artificial a la humana, y hasta se le propone como una inteligencia superior, capaz de sustituir a su predecesora. A esto último es a lo que se le ha llamado singularidad tecnológica, otro problema que ha sido bastante discutido también con relación a las máquinas, su inteligencia y su interacción con el medio y los humanos, pero sobre el cual no nos enfocaremos en este trabajo, como puede advertirse (Cfr. Kurzweil, 2005; Bostrom, 2014; Chalmers, 2010; Vinge, 1993; Good, 1965).

En el caso de las reflexiones que ponen en duda la capacidad de la inteligencia artificial para emular efectivamente la inteligencia humana, la mayoría se centran, ya en la posibilidad de sentir

emociones (Boden, 1996; Tariq et al., 2022), tener consciencia o autoconciencia (Haladjian y Montemayor, 2016; Hildt, 2019; McDermott, 2007; Buttazzo, 2001) y la creatividad (Liu, 2023).

En esas discusiones se pone en tela de juicio si los aspectos conscientes de la emoción podrían explicarse en la Inteligencia Artificial, pues se entiende que la conciencia es a menudo dada por sentado y no se aborda adecuadamente al momento de pensar en estos desarrollos tecnológicos, y se cuestiona tanto si es posible competir en ese sentido con la inteligencia humana como la posibilidad de verificar que un ser inteligente es autoconsciente pues, para algunos, la conciencia se basa en la autoconciencia (McDermott, 2007; Buttazzo, 2001), y es poco probable, se dice en la literatura, que se pueda crear un sistema de inteligencia artificial que tenga realmente algún tipo de experiencia consciente (Haladjian y Montemayor, 2016). De modo que resulta imperativo reconocer que la simulación de la inteligencia humana por parte de entidades mecánicas no conlleva, en sí misma, la atribución de una conciencia, de emociones o de autoconciencia.

Y es que la inteligencia artificial, como tal, se fundamenta en algoritmos y estructuras de datos, los cuales, al no haber emergido de un proceso evolutivo análogo al humano, no se originan de experiencias vinculadas a emociones y una genuina autoconciencia. Aunque la noción precisa de 'conciencia' sigue siendo objeto de debate, es razonable postular que, dado que estos procesos ocurren mediante mecanismos distintos, un estado de conciencia resultante también habría de diferir, especialmente en lo que respecta a la integración de emociones, experiencias percibidas y autoconciencia.

3. PROBLEMAS AXIOLÓGICOS EN LA INTELIGENCIA ARTIFICIAL

Con relación al segundo factor: las cuestiones axiológicas y éticas de la IA, la revisión de la literatura filosófica reciente evidencia que estas cuestiones son de alta relevancia en los debates en torno a ese constructo aun en discusión, pero cuyos frutos son tan tangibles como los productos más tradicionales de la investigación tecnológica.

En esa línea, como ha habido desarrollos de inteligencia artificial enfocada a distintos campos y, en los últimos días hablamos de Inteligencia Artificial Generativa (IAG) -programas informáticos que, al menos en apariencia, son capaces de crear cosas nuevas a partir de peticiones en lenguaje natural de parte de los usuarios- las preocupaciones éticas pueden clasificarse de acuerdo a las distintas áreas del saber que han sido tocadas con mayor incidencia. Estas áreas son la economía, el arte y la cultura, educación, la medicina y la política. Sin embargo, por razones de espacio y enfoque, entre otras, en este trabajo vamos

a dirigir la atención sólo a las dos últimas, con la intención de abordar las otras en ulteriores trabajos con la exhaustividad que ameritan.

3.1. Inteligencia artificial y salud

En el campo de la medicina, la IA ha representado unos avances inusitados que dejarían perplejo a cualquier ignaro en el asunto: con su inigualable capacidad para integrar grandes conjuntos de datos clínicos a modo de "aprendizaje", es capaz de desempeñar importantes roles tanto en diagnóstico como en toma de decisiones clínicas y en medicina personalizada (Rigby, 2019); y como lo muestra Loh (2018), se ha descubierto que los soldados, por ejemplo, frente a un interlocutor virtual generado por algoritmos, muestran una mayor predisposición a desvelar aspectos relacionados con el estrés postraumático; en el ámbito quirúrgico, en 2016 se marcó un hito: un robot, dotado de una inteligencia artificial avanzada, llevó a cabo una intervención en los intestinos de un cerdo con una precisión que superó a la de sus contrapartes humanas. Y, para añadir un grado más de asombro a este panorama, recientemente en China, un robot odontólogo realizó con éxito un implante dental, sin la intervención o supervisión de un profesional humano Loh (2018).

De acuerdo con Hamet y Tremblay (2017), la relación entre IA y medicina puede entenderse en dos vías principales: la virtual y la física. La rama virtual abarca enfoques informáticos que van desde la gestión de información mediante aprendizaje profundo hasta el control de sistemas de gestión de salud, incluyendo registros médicos electrónicos, y la orientación activa de los médicos en sus decisiones de tratamiento, como el caso de aquel robot que entrevista y diagnostica a los soldados tras experiencias de combate. Por otro lado, la rama física se representa mejor a través de robots utilizados para asistir al paciente anciano o al cirujano en funciones. También incorporada en esta rama se encuentran los nanorobots[1] dirigidos, un nuevo sistema de administración de medicamentos que ha demostrado una alta eficiencia.

Pero sea en una vía o en otra, las preocupaciones éticas son latentes. Por ejemplo, Khan et al., (2022) afirman que, siendo la innovación clínica una de las áreas económicamente más prometedoras en la actualidad, la oferta preferencial de equipos clínicos, (como dispositivos de monitoreo cardíaco) a personas jóvenes, que no cumplen realmente con el perfil de usuario previsto inicialmente, representa un problema ético al que hay que atender, puesto que esto está sucediendo porque los jóvenes representan

[1] Dispositivos robóticos de tamaño nanométrico (un nanómetro es una milmillonésima parte de un metro) diseñados para realizar tareas específicas a nivel molecular o celular con aplicaciones potenciales en medicina, fabricación e investigación.

una creciente proporción de ingresos debido a que son menos propensos a sufrir problemas de salud como la fibrilación auricular, por ejemplo.

Además, afirman que el Internet de las Cosas (IoT), que es la posibilidad de que los equipos se comuniquen entre ellos sin intermediación humana compartiendo informaciones relevantes para el desarrollo de la tarea para la que está destinado cada uno de esos equipos, está redefiniendo la idea de una persona sana, pues a través de la comunicación entre dispositivos se manejan informaciones personales e íntimas respecto a la salud de los sujetos que tienen el potencial, no sólo de provocar discriminación al momento de ofrecer coberturas de seguros de salud y otros servicios similares, sino también de aumentar el estigma en torno a aquellas personas crónicamente enfermas o desfavorecidas (Mittelstadt, 2017).

En esa misma línea, Loh (2018) pone el ojo sobre los robots de cuidado, unas máquinas que pueden usarse en hospitales para asistir a pacientes de manera autónoma al estar equipadas de sensores y mecanismos manejados por inteligencia artificial, y recuerda el conocido "dilema del tranvía": una situación en la que un sujeto se encuentra en un tranvía en ruta hacia cinco obreros, y se le presenta la posibilidad de desviar su trayectoria hacia otra vía donde se encuentra un solo trabajador. La cuestión filosófica que surge es si es éticamente defendible sacrificar a un individuo en pos del bienestar de cinco. Este dilema adquiere especial relevancia en el contexto de estos robots e IA de uso médico, pues no es descabellado pensar en un escenario real donde un robot deba decidir salvar a un paciente a expensas de otro. Y aquí, estaríamos frente a un problema de índole moral: ¿quién está en posición de tomar tal decisión? ¿El artífice del software o el médico a cargo?

Del mismo modo, Sullivan y Schweikar (2019) describen una situación en la que un sistema de inteligencia artificial del tipo "black-box" (caja negra[2]) se utiliza para asistir en la detección del cáncer de mama mediante datos de mamografía. Si este sistema sugiere un diagnóstico erróneo, podría resultar en una lesión o daño para el paciente, pero las leyes públicas, en su estado actual, no contemplan forma alguna de proceder en estos casos, lo que deja al agraviado en una situación de desamparo.

Finalmente, tal y como lo presentan Cobianchi et al. (2022), en gran parte de la literatura académica que puede hallarse en las bases de datos se llama la atención sobre el peligro de los casos de discriminación relacionados con el género y la etnia debido a datos sesgados. Esto así porque, de manera

[2] El término "caja negra" en el contexto de la inteligencia artificial se refiere a sistemas o modelos cuyo funcionamiento interno es incomprensible o no transparente para los humanos. Aunque podemos observar las entradas y salidas de estos sistemas, no podemos entender fácilmente cómo llegan a sus decisiones o conclusiones específicas.

específica, el uso de algoritmos de Redes Neuronales Profundas[3] provoca el fenómeno de caja negra y tiene la potencialidad de heredar sesgos en los datos que, a posteriori, afectarían la toma de decisiones perjudicando a sujetos en función de uno o varios atributos (como género o etnia) que no están relacionados con los atributos previstos como necesarios para esa toma decisión.

Este último ejemplo nos deja en condición de abordar las discusiones que se han suscitado en el campo de la política con relación a la inteligencia artificial.

3.2. Inteligencia artificial y política

Como se ha dicho, también en la política la IA ha desplegado notables progresos, específicamente en lo que respecta a políticas de control y mecanismos operativos. Sin embargo, nos encontramos ante una importante ausencia de un marco ético-decisional coherente y, lo que es más grave, de un cuerpo normativo que aborde los dilemas éticos inherentes a estas prácticas. Esta situación, lejos de ser trivial, obstaculiza representa un foco esencial al cual mirar cuando desde la política lo que se intenta es construir una sociedad justa, administrada transparentemente y que garantice seguridad y tranquilidad a sus ciudadanos.

Partiendo de que es un hecho que muchas organizaciones están realizando altas inversiones en soluciones de inteligencia artificial (IA), por el supuesto potencial de la IA para mejorar el bienestar organizacional y humano, Telkamp y Anderson (2022) plantean que, aunque en términos generales las personas quieren que tanto las empresas que usan IA como los propios sistemas de IA se comporten éticamente, existe una dificultad a superar ante este aparente consenso: y es que, a eso que se le llama comportamiento ético, significa cosas diferentes para diferentes personas. A este problema es lo que se ha denominado el problema del pluralismo axiológico. Y si bien es cierto que se han propuesto algunos marcos éticos para la IA, algunos autores que los han examinado concluyen que sus enfoques no abordan adecuadamente los mecanismos por los cuales los sujetos realizan las evaluaciones éticas de los sistemas de IA o no integran al menos los desacuerdos fundamentales sobre lo que es y no es un comportamiento ético (Telkamp y Anderson, 2022).

De forma general, en términos políticos, los principales desafíos que llaman la atención de los investigadores se ven representados en lo que respecta a la desinformación, la polarización, la vulnerabilidad de los sistemas de información y la invasión de la privacidad.

3 Algoritmos inspirados en el cerebro humano, con múltiples capas que procesan y modelan datos complejos. Son ampliamente utilizados en IA para tareas como reconocimiento de imágenes y procesamiento del lenguaje natural.

Las redes sociales, los bots[4] y otras plataformas digitales, así como los denominados *trolls*[5] en línea, están siendo potenciados por la inteligencia artificial para difundir información falsa o engañosa (*fake news*), y esto suele ser utilizado para manipular discusiones en línea y amplificar mensajes específicos que terminan manipulando la percepción pública y erosionando la confianza en las instituciones, que son consecuencias directas de la desinformación y la polarización (Cfr. López et al., 2023; Serra Cristóbal, 2023).

Adicionalmente, los algoritmos de las plataformas de medios sociales, como están diseñados para maximizar la interacción del usuario, tienden a mostrar contenidos que refuerzan creencias preexistentes, creando lo que se conoce como "cámaras de eco"[6] que masifican y maximizan la desinformación (Cfr. Rodríguez, 2022; Pietropaoli, 2022).

En cuanto a la vulnerabilidad de los sistemas de información, esta se refiere de manera muy directa a los riesgos de ciberataques: intentos maliciosos de un individuo o una organización para acceder, alterar, robar o destruir información en sistemas informáticos o redes, o para interrumpir los servicios normales de estos sistemas. La preocupación política principal radica en el hecho de que los sistemas electorales, así como otras infraestructuras críticas suelen ser vulnerables a estos ciberataques y, en la actualidad, esa vulnerabilidad se ve potenciada por la existencia de técnicas avanzadas de inteligencia artificial (Cfr. Navío Marco, 2023; Huapaya Noriega, 2022).

Finalmente, la invasión de la privacidad por medio de la IA adquiere carácter político en tanto las tecnologías de vigilancia y recolección de datos, respaldadas por los denominados algoritmos de aprendizaje automático, pueden ser empleadas para monitorear a ciudadanos y opositores políticos, lo que puede traer como consecuencia restricciones de la libertad de expresión y violaciones de derechos fundamentales por parte de gobiernos con inclinación al totalitarismo y la dictadura (Momen, 2023).

En esta misma línea de preocupación se enmarcan las denominadas ciudades inteligentes: un territorio que, combinando innovación, infraestructura digital y gestión pública respaldada por tecnologías de comunicación avanzadas, ameritan de la implementación de sistemas de vigilancia e intercomunicación entre los subsistemas de gobernanza para lograr la perseguida eficiencia en la gestión

4 Programa informático diseñado para realizar tareas automáticas en línea que pueden variar desde responder preguntas frecuentes hasta realizar acciones repetitivas en sitios web o redes sociales, y que pueden ser utilizados para una variedad de propósitos, tanto legítimos como maliciosos.

5 Persona que publica mensajes provocativos, ofensivos o fuera de tema en comunidades en línea, como foros, chats o redes sociales, con la intención de molestar, provocar o desencadenar reacciones negativas en otros usuarios.

6 Fenómeno que se da cuando individuos reciben información que refuerza sus creencias existentes, principalmente en redes sociales, limitando la exposición a opiniones divergentes (Sunstein, 2001).

pública, y esto incluye un monitoreo constante de los ciudadanos en las áreas públicas, así como sobre los accesos a los servicios públicos.

Esta idea de ciudades inteligentes, en cuya lista ya se encuentran algunas de las grandes ciudades del mundo, como puede verse implica una fundamentación en servicios y sistemas que hace uso de la inteligencia artificial en su lógica operativa. Y el hecho de que el buen desarrollo de esas funciones esté basado en recopilación, intercambio y procesamiento de grandes cantidades de datos e información de los ciudadanos, puede significar tanto un viaje emocionante para descubrir las nuevas oportunidades como un camino incierto hacia consecuencias siniestras cuando los datos de los usuarios se usan indebidamente (Cfr. Momen, 2023). Esto así porque los datos se comparten entre múltiples partes interesadas, y esto pone en riesgo la privacidad de las personas.

4. CONCLUSIONES

La inteligencia artificial (IA), en su intersección con la filosofía, nos invita a reflexionar sobre la naturaleza de la inteligencia, la conciencia y la ética, y es importante continuar este diálogo interdisciplinario a fin de comprender y prever mejor el futuro tecnológico que se avecina. En su esencia, se trata de una herramienta creada por el ser humano, sin embargo, su capacidad para procesar información, y simular aprendizaje y adaptación, la coloca en una posición única en comparación con otras herramientas tecnológicas.

En cuanto a la definición y comprensión de la inteligencia, tanto la biológica como la artificial, sigue siendo un tema de discusión. Como se ha visto, a pesar de que la IA puede simular procesos cognitivos humanos, el hecho de que no posea conciencia, emociones o autoconciencia de la misma manera que los seres humanos, deja descartado, en términos lógicos, el debate sobre si la inteligencia artificial -donde "inteligencia" describe procesos mecánicos y digitales- se puede equiparar con la inteligencia humana, toda vez que en esta última se involucran esos aspectos que, por lo pronto, no son vislumbrables en el horizonte de posibilidades de aquella.

No obstante, en el ámbito de la salud, la IA ha demostrado ser una herramienta valiosa, con potencial para mejorar la precisión en el diagnóstico y el tratamiento. Pero su uso también plantea preocupaciones éticas, especialmente en relación con la toma de decisiones médicas y la privacidad del paciente.

Igualmente, en el ámbito político, la IA tiene el potencial de influir en la opinión pública, ya sea a través de la difusión de desinformación o la creación de cámaras de eco. Adicionalmente, la creciente

dependencia de la tecnología en la infraestructura crítica estatal plantea preocupaciones sobre la vulnerabilidad a ciberataques, y la idea de "ciudades inteligentes", aunque prometedora, plantea preocupaciones sobre la privacidad y la seguridad de los datos de ciudadanos y organizaciones.

Visto lo anterior, puede concluirse que la inteligencia artificial ha impactado múltiples campos del saber humano, que su desarrollo plantea importantes cuestionamientos éticos y que estas cuestiones éticas relacionadas con la IA son de suma importancia, no sólo en áreas como la salud y la política, sobre las que nos hemos detenido acá, sino que ese impacto y preocupación es extensible a campos como la educación, la economía, el arte y la cultura. Por ello, resulta imperativo abordar estos temas con rigor y cautela a los fines de promover un desarrollo y aplicación responsables de la tecnología.

Finalmente, esa reflexión, y el diálogo sobre la IA, entre otras cosas, deben orientarse hacia la construcción de marcos éticos y normativos que puedan guiar el desarrollo y la aplicación de la IA que propicien que estos avances beneficien a la sociedad en su conjunto, o que, cuando menos, se puedan establecer con claridad los principios que guían su producción, desarrollo, implementación, distribución y uso.

5. REFERENCIAS

Adams, R., Víctor, M., & Ropper, A. (Eds). (1997). Chapter 21. Dementia and the Amnesic (Korsakoff) Syndrome. En: Principles of Neurology. 6th Edition. McGraw Hill. New York.

Alonso-Fernández, F. (1996). El talento creador. Rasgos y perfiles del genio. Ediciones Temas de Hoy. Madrid.

Bezlepkin, E. A., & Zaykova, A. S. (2021). Neurofilosofía, filosofía de las neurociencias y filosofía de la inteligencia artificial: el problema de la distinción. Философские науки, 64(1), 71-87. https://doi.org/10.21146/2413-9084-2021-64-1-71-87

Buttazzo, G. (2001). Artificial consciousness: Utopia or real possibility? Computer, 34(7), 24-30. https://doi.org/10.1109/2.933500.

Cobianchi, L., Verde, J. M., Loftus, T. J., Piccolo, D., Dal Mas, F., Mascagni, P., ... & Kaafarani, H. M. (2022). Artificial Intelligence and Surgery: Ethical Dilemmas and Open Issues. J Am Coll Surg, 235(2), 268-275. https://doi.org/10.1097/XCS.0000000000000242

Descartes, R. (1664). Tratado del hombre. Traducción: Guillermo Quintas. Ilustraciones: Gérard van Gutschoven y Louis de la Forge. Ilustración de cubierta: Dino Valls, De Profundis, 1989. Editor digital: RLull.

Di Liscia, D., & Sylla, E. (Eds.). (2022). Quantifying Aristotle. The Impact, Spread and Decline of the Calculatores Tradition. Leiden: Brill.

Friedman, M. (2023). Leibniz and the Stocking Frame: Computation, Weaving and Knitting in the 17th Century. Minds & Machines. https://doi.org/10.1007/s11023-023-09623-3

Gardner, H. (1997). Estructuras de la Mente: La teoría de las Inteligencias Múltiples. Fondo de Cultura Económica. Bogotá.

Ghallab, M., Nau, D., & Traverso, P. (2004). Automated Planning: Theory & Practice. Elsevier.

Haladjian, H. H., & Montemayor, C. (2016). Artificial consciousness and the consciousness-attention dissociation. Consciousness and Cognition, 45, 210-225. https://doi.org/10.1016/j.concog.2016.08.011

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. Metabolism, 69(Supplement), S36-S40. https://doi.org/10.1016/j.metabol.2017.01.011

Haugeland, J. (1985). Artificial Intelligence: The Very Idea. MIT Press.

Huapaya Noriega, D. A. (2022). Aportes para la política exterior peruana en materia de amenazas híbridas en el ciberespacio [Tesis de maestría, Academia Diplomática del Perú

Javier Pérez de Cuéllar]. Repositorio Institucional - ADP. http://repositorio.adp.edu.peADP/209

Khan, S., Chanchal, D. K., Singh, G., Gupta, S., & Porwal, P. (2022). Artificial Intelligence as an Essential Tool for the Development of Medicine Requirement. Research Journal of Medicine and Pharmacy, 1(1). https://embarpublishers.com/assets/articles/1666164307.pdf

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, 1-11. https://doi.org/10.5209/aris.83808.

Loh, E. (2018). Medicine and the rise of the robots: a qualitative review of recent advances of artificial intelligence in health. BMJ Leader, 2(2), 59-63. https://doi.org/10.1136/leader-2018-000071

López, P. C., Maldonado, A. M., & Ribeiro, V. (2023). La desinformación en las democracias de América Latina y de la península ibérica: De las redes sociales a la inteligencia artificial (2015-2022). Uru: Revista Latinoamericana de Comunicación, 8(5), 69-89. https://doi.org/10.32719/26312514.2023.8.5

Margaret Boden. (1996). Artificial Intelligence. Academic Press. California.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. The MIT Press.

McCarthy, J., & Hayes, P. J. (1981). Some Philosophical Problems from the Standpoint of Artificial Intelligence. En B. L. Webber & N. J. Nilsson (Eds.), Readings in Artificial Intelligence (pp. 431-450). Morgan Kaufmann. https://doi.org/10.1016/B978-0-934613-03-3.50033-7

McDermott, D. (2007). Artificial Intelligence and Consciousness. En P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), The Cambridge Handbook of Consciousness (pp. 117-150). Cambridge University Press.

Merejo, A. (2023a). Filosofía para tiempos transidos y cibernéticos. Editorial Santuario.

Merejo, A. (2023b). Cibermundo transido: Enredo gris de pospandemia, guerra y ciberguerra. Editorial Santuario

Mittelstadt, B. (2017). Ethics of the healthrelated internet of things: a narrative review. Ethics and Information Technology, 19(3), 157-175. https://doi.org/10.1007/s10676-017-9426-4

Momen, N. (2023). Privacy and Ethics in a Smart City: Towards Attaining Digital Sovereignty. In: Ahmed, M., Haskell-Dowland, P. (eds) Cybersecurity for Smart Cities. Advanced

Sciences and Technologies for Security Applications. Springer, Cham. https://doi-org.ehu.idm.oclc.org/10.1007/978-3-031-24946-4_4

Navío Marco, J. (2023). La economía digital y la digitalización de las pymes: Un enfoque desde la Unión Europea. En J. Navío Marco (Coord.), Economía Digital en la Unión Europea: Apoyando a las Pymes (pp. 3-90). Editorial Sanz y Torres. https://empresas.blogthinkbig.com/wp-content/uploads/2023/04/Economia-Digital-en-la-Union-Europea.pdf

Pietropaoli, S. (2022). De ciudadano a usuario: Capitalismo, democracia y revolución digital. Revista Jurídica de Asturias, (45), 13-22. https://reunido.uniovi.es/index.php/RJA/article/view/18983/15332

Rigby, M. J. (2019). Ethical Dimensions of Using Artificial Intelligence in Health Care. AMA Journal of Ethics, 21(2), 121-124. https://doi.org/10.1001/amajethics.2019.121

Rodríguez, J. M. (2022). Violencia estructural y las tecnologías de la información. Trayectorias Humanas Trascontinentales, (14). https://doi.org/10.25965/trahs.4850

Russell, S. J., & Norvig, P. (2010). Artificial Intelligence: A Modern Approach. Prentice Hall.

Serra Cristóbal, R. (2023). Noticias falsas (fake news) y derecho a recibir información veraz. Dónde se fundamenta la posibilidad de controlar la desinformación y cómo hacerlo. Revista de Derecho Político, (116), 13–46. https://doi.org/10.5944/rdp.116.2023.37147

Sullivan, H. R., & Schweikart, S. J. (2019). Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI? AMA J Ethics, 21(2), 160-166. https://doi.org/10.1001/amajethics.2019.160

Sunstein, C. R. (2001). Republic.com. Princeton University Press.

Tariq, S., Iftikhar, A., Chaudhary, P., & Khurshid, K. (2022). Examining Some Serious Challenges and Possibility of AI Emulating Human Emotions, Consciousness, Understanding and 'Self'. Journal of NeuroPhilosophy, 1(1). https://doi.org/10.5281/zenodo.6637757.

Telkamp, J. B., & Anderson, M. H. (2022). The Implications of Diverse Human Moral Foundations for Assessing the Ethicality of Artificial Intelligence. Journal of Business Ethics: JBE, 178(4), 961-976. https://doi.org/10.1007/s10551-022-05057-6

Turing, A. (1950). Computing Machinery and Intelligence. Mind, 59(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

Winograd, T., & Flores, F. (1986). Understanding Computers and Cognition: A New Foundation for Design. Addison-Wesley.

CÓMO CITAR:

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