Escritos / Medellín-Colombia / Vol. 33, N. 71 / pp. 1-19
julio-diciembre, 2025 / ISSN-e: 2390-0032

Cómo citar en Chicago: León, Eduardo Alberto. "Kant y la cuestión del juicio moral en la era de la inteligencia artificial". *Escritos,* Vol. 33, no. 70 (2025): 1-19. https://doi.org/10.18566/escr.v33n71.a02

Fecha de recepción: 27 de septiembre, 2025
Fecha de aceptación: 7 de noviembre, 2025

Kant y la cuestión del juicio moral en la era de la inteligencia artificial

Kant and the Question of Moral Judgment in the Age of Artificial Intelligence

Eduardo Alberto León[1]

1 Doctor en Sociología por FLACSO-Ecuador, Máster en Educación por la UIDE (Ecuador), Máster en Filosofía y Pensamiento Social por FLACSO-Ecuador y Licenciado en Filosofía por Ottawa University. Facultad Latinoamericana de Ciencias Sociales (FLACSO). Correo electrónico: alberto3026@yahoo.es.

RESUMEN
Este artículo analiza la posibilidad de atribuir juicio moral a la inteligencia artificial (IA) desde la perspectiva kantiana. Se sostiene que, aunque los sistemas de IA pueden simular conductas éticas mediante algoritmos y reglas preprogramadas, carecen de libertad trascendental y de una voluntad autónoma, condiciones esenciales en la moralidad kantiana. El estudio integra conceptos fundamentales de la filosofía práctica –como el deber y el imperativo categórico– con experimentos exploratorios que ilustran las respuestas de sistemas de IA en dilemas morales. Los hallazgos muestran que las acciones de las máquinas son amorales, pues se limitan a fines externos determinados por sus programadores. Se concluye que la ética kantiana constituye un marco crítico para subrayar los límites del juicio moral artificial y resalta la necesidad de mantener un enfoque humano-céntrico en el desarrollo de estas tecnologías.

Palabras clave: Ética kantiana, Juicio moral, Inteligencia artificial, Autonomía, Imperativo categórico.

ABSTRACT
This article analyzes the possibility of attributing moral agency to artificial intelligence (AI) from a Kantian perspective. It argues that although AI systems can simulate ethical behavior through preprogrammed algorithms and rules, they lack transcendental freedom and autonomous will, which are essential conditions for Kantian morality. The study combines key concepts of practical philosophy –such as duty and the categorical imperative– with exploratory experiments illustrating AI responses to moral dilemmas. Findings show that machine actions are amoral, as they are limited to external ends determined by human programmers. The article concludes that Kantian ethics provides a critical framework to highlight the limits of artificial moral agency and emphasizes the need for a human-centered approach in AI development.

Keywords: Kantian ethics, Moral judgment, Artificial intelligence, Autonomy, Categorical imperative.

Introducción

La inteligencia artificial (IA) fue concebida originalmente como una herramienta para amplificar las capacidades humanas, un concepto que Nick Bostrom describe como "inteligencia aumentada".[2] En este contexto, alcanzar una IA de nivel humano, mediante la emulación del cerebro, depende de avances tecnológicos significativos, como "la exploración de alto rendimiento".[3] No obstante, el propósito de la IA ha evolucionado con el tiempo, pasando de ser una extensión de la inteligencia humana a la creación de agentes autónomos capaces de tomar decisiones éticas sin supervisión directa. Este cambio plantea preguntas fundamentales sobre la posibilidad y deseabilidad de otorgar a las máquinas responsabilidades morales.

Esta transición hacia una IA éticamente autónoma no estuvo prevista por los pioneros del campo. Como señala Bostrom, "los pioneros de la IA, en su mayoría, no contemplaron la posibilidad de que su empresa pudiera implicar ningún riesgo. No hablaron –ni pensaron seriamente– sobre ningún problema de seguridad o cuestión ética relacionada con la creación de mentes artificiales".[4] Esta omisión inicial ha generado un interés creciente en desarrollar sistemas que no solo sean funcionalmente competentes, sino también capaces de alinearse con principios éticos humanos. Lograr que las máquinas tomen decisiones moralmente correctas requiere, por tanto, la formulación de una ética específica para la IA, un desafío que divide a los expertos en dos perspectivas opuestas. Por un lado, los defensores de la IA fuerte sostienen que es posible crear agentes artificiales con autonomía ética plena, capaces de razonar y actuar como agentes morales independientes. Por otro lado, los partidarios de la IA débil argumentan que las máquinas, por sofisticadas que sean, no pueden poseer la capacidad de juicio moral propia de los seres humanos.

Tradicionalmente, la ética se ha asociado con la capacidad de razonar y evaluar las consecuencias de las acciones. En este sentido, Espinoza Freire y Calva Nagua definen la conducta ética como el resultado de "juicios de comportamiento, al poner en práctica con su accionar el sistema de valores éticos aprendidos a lo largo de la vida y las normas establecidas por la sociedad o grupo al cual pertenece".[5] Dado que esta definición implica un nivel de consciencia y racionalidad, atribuir cualidades éticas a sistemas artificiales resulta problemático, ya que históricamente solo los seres humanos han sido considerados agentes morales.

A pesar de estas reservas, la integración de sistemas de IA en la vida cotidiana es innegable y sus decisiones tienen un impacto significativo en la sociedad, tanto positivo como negativo. La posibilidad de que estos sistemas actúen como agentes morales independientes plantea una cuestión crítica: ¿cómo podemos garantizar que las acciones de la inteligencia artificial se alineen con los valores éticos humanos? Responder a esta pregunta requiere no solo avances tecnológicos, sino también un marco ético robusto que permita maximizar los beneficios de la IA y mitigar sus riesgos. En este sentido, evaluar el desempeño ético de los sistemas inteligentes se convierte en un paso esencial para orientar el desarrollo de esta tecnología hacia un impacto positivo y sostenible.

2 Nick Bostrom, *Superinteligencia: Caminos, peligros, estrategias* (Oxford: Oxford University Press, 2016).

3 Bostrom, *Superinteligencia,* 75.

4 Ibid., 30.

5 Eudaldo Enrique Espinoza Freire y Daniel Xavier Calva Nagua, "La ética en las investigaciones educativas", *Revista Universidad y Sociedad*, Vol. 12, no. 4 (2020): 335.

La posibilidad de atribuir un juicio moral a sistemas de IA plantea interrogantes éticas sobre la consistencia de sus decisiones. En este contexto, Isaac Asimov[6] propuso tres leyes de la robótica como un marco ético inicial para robots[7]. Sin embargo, estas leyes, aunque visionarias, resultan insuficientes para guiar el comportamiento de los sistemas autónomos actuales, que desempeñan un papel central en nuestra vida cotidiana, desde asistentes virtuales hasta tecnologías predictivas basadas en sensores.

Aunque se espera que en el futuro pueda desarrollarse una inteligencia artificial general (IAG) con capacidades superiores a las humanas, los avances actuales aún están lejos de ese objetivo. No obstante, el impacto de la IA es significativo y un mal funcionamiento podría generar consecuencias graves para la sociedad. Como advirtió Elon Musk, "con la inteligencia artificial, estamos invocando al demonio",[8] una metáfora que resalta la urgencia de diseñar principios éticos desde las primeras fases de desarrollo para garantizar que la IA contribuya positivamente al bienestar humano.

Para abordar esta problemática, es crucial determinar si las máquinas pueden considerarse capaces de juicio moral y cómo este difiere del juicio humano. La ética kantiana define al agente moral como un ser capaz de actuar según principios racionales universales, pero lo que verdaderamente caracteriza a la moralidad kantiana es la capacidad de formarse un juicio sobre lo que debe hacerse, es decir, de deliberar y aplicar principios universales a situaciones concretas. Evaluar los sistemas de IA desde la perspectiva del juicio moral permite esclarecer si pueden asumir responsabilidades éticas de manera análoga a los seres humanos, sentando las bases para un desarrollo tecnológico alineado con valores éticos fundamentales. La evaluación de la capacidad de los sistemas de IA para ejercer juicio moral requiere situarlos dentro de nuestro marco ético, predominantemente antropocéntrico. La ética kantiana, que centra la moralidad en la razón humana, ofrece una perspectiva útil para este análisis, aunque choca con visiones posthumanistas que cuestionan la exclusividad del juicio moral humano.[9] Immanuel Kant[10] no abordó explícitamente la posibilidad de un juicio moral no humano, lo que nos invita a extender sus conceptos para explorar si una "máquina kantiana" podría ser considerada capaz de emitir un juicio moral auténtico.

Desde la perspectiva kantiana, el juicio moral se fundamenta en los conceptos de deber, imperativo categórico y libertad de la voluntad. Kant sostenía que los seres humanos, en tanto dotados de razón, son capaces de ejercer un juicio moral imparcial, pero dicha imparcialidad, así como la ausencia de afectividad, solo es posible en un plano puramente racional, en la medida en que la humanidad sea tomada como fin en sí misma y nunca como medio.[11] En contraste, los sistemas de IA, aunque pueden simular comportamientos éticos mediante procesos de racionalidad algorítmica, dependen en última

6 Isaac Asimov, *Yo, robot*, trad. L. G. Prado (Barcelona: Edhasa, 2009).

7 1) Un robot no puede dañar a un ser humano o, por inacción, permitir que sufra daño; 2) un robot debe obedecer las órdenes dadas por seres humanos, salvo que contradigan la primera ley, y 3) un robot debe proteger su propia existencia, siempre que no entre en conflicto con las dos primeras leyes. Asimov, *Yo, robot.*

8 Matt McFarland, "Elon Musk: 'with Artificial Intelligence We Are Summoning the Demon'", *The Washington Post*, October 24, 2014, https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/.

9 José Ignacio Murillo, "¿Necesita ética el posthumanismo? La normatividad de una naturaleza abierta", *Cuadernos de Bioética*, Vol. 25, no. 3 (2014).

10 Immanuel Kant (2007).

11 Kant, 2008.

instancia de las decisiones y valores de sus programadores humanos, lo que limita su autonomía moral. Podría argumentarse que una ética basada en reglas, como la kantiana, resulta particularmente apropiada para aplicarse en sistemas artificiales, dado que establece deberes y principios de acción que, en su mayoría, son computacionalmente tratables. No obstante, la falta de una voluntad consciente en la IA plantea serias dudas respecto a su capacidad para encarnar un juicio moral genuino, tal como la concibe Kant. Como advierte Zetina Gutiérrez, "una cosa es segura, desde la perspectiva de la filosofía kantiana, delegar la tarea de pensar los números a la IA no traerá mejores respuestas que las que la humanidad ya había vislumbrado".[12] Esta reflexión subraya los límites epistemológicos y morales de atribuir a las máquinas un papel análogo al del sujeto racional humano en el marco de la ética kantiana.

En consecuencia, este artículo se propone analizar, desde la ética kantiana, si los sistemas de IA pueden considerarse agentes morales. Para ello, se presentarán tanto los fundamentos teóricos como experimentos exploratorios con IA, a fin de mostrar que, aunque los sistemas actuales pueden simular comportamientos éticos, carecen de las condiciones trascendentales que definen la moralidad auténtica.

Kant y la fundamentación del juicio moral

Habitualmente, se designa como "agente moral" a todo ser humano que demuestra poseer una conciencia moral. Estos individuos tienen la capacidad de evaluar si una acción es legítima o no, considerando su impacto en el bienestar de la sociedad humana. En este contexto, Kant argumentó, en la *Crítica de la razón pura*, que únicamente los seres humanos pueden comprender la relación causal, a menudo condicionada por necesidades patológicas, entre las acciones y sus consecuencias, y que la "libertad trascendental" es una característica inherente a los seres racionales. Así lo explica Kant:

> Merece especial atención el hecho de que la idea trascendental de la libertad sirva de fundamento al concepto práctico de ésta y que aquélla represente la verdadera dificultad que ha implicado desde siempre la cuestión acerca de la posibilidad de esa libertad. En su sentido práctico, la libertad es la independencia de la voluntad respecto de la imposición de los impulsos de la sensibilidad. En efecto, una voluntad es sensible en la medida en que se halla patológicamente afectada (por móviles de sensibilidad). Se llama animal (*arbitrium brutum*) si puede imponerse patológicamente. La voluntad humana es arbitrium sensitivum, pero no *brutum*, sino liberum, ya que la sensibilidad no determina su acción de modo necesario, sino que el hombre goza de la capacidad de determinarse espontáneamente a sí mismo con independencia de la imposición de los impulsos sensitivos.[13]

Kant señala que la idea trascendental de libertad –es decir, la libertad concebida más allá de la experiencia sensible– es el fundamento necesario para entender la libertad práctica, que implica la capacidad de la voluntad para actuar independientemente de los impulsos sensibles. A diferencia de los animales, cuya voluntad está determinada por estímulos externos, el ser humano posee una libertad que le permite autodeterminarse racionalmente. Así, la verdadera dificultad no está tanto en definir la libertad práctica,

12 Humberto Zetina Gutiérrez, "Kant y los límites de la inteligencia artificial", *Universitaria*, Vol. 7, no. 48 (2024): 38, https://revistauniversitaria.uaemex.mx/article/view/22699.

13 Immanuel Kant, *Crítica de la razón pura*, trad. Pedro Ribas (Madrid: Taurus, 2005), 340.

sino en justificar cómo es posible esta autonomía frente a la influencia natural de los sentidos. Asimismo, Kant argumenta:

> Es fácil de ver que, si toda la causalidad que hay en el mundo sensible fuese meramente naturaleza, todo acontecimiento estaría temporalmente determinado por otro según leyes necesarias. Por tanto, teniendo en cuenta que los fenómenos deberían, en la medida en que determinan la voluntad, hacer necesaria cada acción como resultado natural de los mismos, la supresión de la libertad trascendental significaría, a la vez, la destrucción de toda libertad práctica. Esta se supone tan sólo un carácter tan sólo dinámico de la libertad en el caso de que no haya sucedido, y, que por consiguiente, la causa de ese algo no puede tener en el efecto del mecanismo un carácter tan determinante, que no pueda haber en nuestra voluntad una causalidad capaz de producir –con independencia de esas causas naturales e incluso contra su poder y su influjo– algo que en el orden temporal se halle determinado según leyes empíricas, una serie que sea, por tanto, capaz de iniciar con entera espontaneidad una serie de acontecimientos.[14]

La "libertad trascendental" constituye el pilar esencial de la moralidad. Kant propuso una forma distinta de causalidad que opera de manera paralela a la causalidad científica, a la que denominó "causalidad trascendental". Toda acción guiada por esta causalidad se libera de las cadenas causales del mundo sensible y nuestro "libre albedrío" se manifiesta como una expresión de dicha "causalidad trascendental". En este punto, Kant introduce el concepto de elección como una capacidad que nos permite trascender la determinación causal científica y facilita la realización de actos morales. Llamó a los seres humanos seres racionales, dado que están sujetos simultáneamente a leyes fenoménicas y trascendentales.

No obstante, la capacidad de "elección" nos permite resistir las inclinaciones del mundo sensible y actuar conforme a nuestra conciencia moral. Esta facultad nos distingue de otros seres condicionados por impulsos, como los animales. Kant sostuvo que los juicios morales requieren la libertad de elegir. Por ende, aquellos que están exclusivamente sometidos a una de estas causalidades no pueden ser considerados juicios morales.[15] Los animales, al estar sujetos únicamente a la causalidad sensible y carecer de la capacidad racional necesaria para el discernimiento moral, no poseen juicio moral.

> Una voluntad perfectamente buena hallaríase, pues, igualmente bajo leyes objetivas (del bien); pero no podría representarse como constreñida por ellas a las acciones conformes a la ley, porque por sí misma, según su constitución subjetiva, podría ser determinada por la sola representación del bien. De aquí que para la voluntad divina y, en general, para una voluntad santa, no valgan los imperativos: el "debe ser" no tiene aquí lugar adecuado, porque el querer ya de suyo coincide necesariamente con la ley. Por eso son los imperativos solamente fórmulas para expresar la relación entre las leyes objetivas del querer en general y la imperfección subjetiva de la voluntad de tal o cual ser racional; verbigracia, de la voluntad humana.[16]

En este contexto, Kant explicó que, si una persona estuviera exclusivamente sometida a la causalidad trascendental, no tendría la posibilidad de actuar de manera diferente. A estos seres los llamó "voluntad

14 Kant, *Crítica de la razón pura*, 340.

15 Kant, *Crítica de la razón pura*.

16 Immanuel Kant, *Fundamentación de la metafísica de las costumbres*, trad. M. García Morente (Madrid: Espasa-Calpe, 2007), 28-29.

santa",[17] una voluntad "cuyas máximas concuerdan necesariamente con las leyes de la autonomía" y que representa "una voluntad absolutamente buena".[18] Sus acciones están completamente alineadas con la razón moral, sin espacio para elegir actuar en contra de ella. El concepto de deber cobra sentido únicamente para los seres humanos, quienes enfrentan la posibilidad de actuar en contra de su "razón". A pesar de esto, optan por guiarse por el pensamiento racional para actuar de manera "moral". Fue en este marco que Kant introdujo la idea del "imperativo", como una guía para la acción moral.

Según Kant, todas las normas morales deben alinearse con el imperativo categórico, una regla universal que se distingue claramente de los imperativos hipotéticos. En sus palabras, el imperativo categórico es aquel que "representase una acción por sí misma, sin referencia a ningún otro fin, como objetivamente necesaria".[19] A diferencia de los imperativos hipotéticos, que dependen de condiciones externas como la búsqueda de "felicidad" o "placer" para ser aplicables, el imperativo categórico encuentra su justificación en sí mismo. Su carácter universal y evidente lo hace intrínsecamente necesario, independiente de cualquier circunstancia particular. Por el contrario, los imperativos hipotéticos están condicionados y carecen de esta universalidad. Para Kant, nuestra libertad de elección debe guiarse exclusivamente por el imperativo categórico si queremos actuar de manera verdaderamente moral. Según él,

> La necesidad natural era una heteronomía de las causas eficientes; pues todo efecto no era posible sino según la ley de que alguna otra cosa determine a la causalidad la causa eficiente. ¿Qué puede ser, pues, la libertad de la voluntad sino autonomía, esto es, propiedad de la voluntad de ser una ley para sí misma? Pero la proposición: "la voluntad es, en todas las acciones, una ley de sí misma", caracteriza tan sólo el principio de no obrar según ninguna otra máxima que la que pueda ser objeto de sí misma, como ley universal. Ésta es justamente la fórmula del imperativo categórico y el principio de la moralidad; así, pues, voluntad libre y voluntad sometida a leyes morales son una y la misma cosa.[20]

Según Kant, en la naturaleza todo efecto tiene una causa externa, un fenómeno que él denomina heteronomía, donde nada ocurre por sí mismo, sino que depende necesariamente de algo más. En contraste, la libertad de la voluntad no consiste en actuar sin reglas, sino en la autonomía, es decir, la capacidad de la voluntad para darse su propia ley. Esto implica obrar conforme a principios o máximas que puedan valer como leyes universales para todos, lo que Kant identifica como el imperativo categórico, fundamento de la moralidad. Por ello, para Kant, una voluntad libre y una voluntad moral son equivalentes, ya que ambas se guían, de manera autónoma, por una ley universal que la razón reconoce como válida. En sus propias palabras, "Felicidad es la satisfacción de todas nuestras inclinaciones (tanto extensive, atendiendo a su variedad, como intensive, respecto de su grado, como también potencie, en relación con su duración). La ley práctica derivada del motivo de la felicidad la llamo pragmática (regla de prudencia). En cambio, la ley, si es que existe, que no posee otro motivo que la dignidad de ser feliz

17 La noción de "voluntad santa" en la filosofía de Kant se refiere a una voluntad que actúa siempre en perfecta conformidad con la razón y la ley moral, sin posibilidad de desviarse debido a inclinaciones o deseos sensibles. A diferencia de los seres humanos, que enfrentan la tensión entre la razón y las tentaciones del mundo sensible, la voluntad santa carece de esta dualidad, ya que está exclusivamente sometida a la causalidad trascendental.

18 Immanuel Kant, *Crítica de la razón práctica*, trad. J. M. Palacios (Madrid: Alianza Editorial, 2008), 439.

19 Kant, *Fundamentación de la metafísica*, 29.

20 Ibid., 59-60.

la llamo ley moral (ley ética). La primera nos aconseja qué hay que hacer si queremos participar de la felicidad. La segunda nos prescribe cómo debemos comportarnos si queremos ser dignos de ella".[21]

La ley pragmática opera bajo lo que Kant denomina principios prácticos materiales, todos los cuales derivan del principio general del amor propio o la búsqueda de la propia felicidad. Esta búsqueda de la "conciencia del carácter agradable de la vida" representa precisamente lo que él considera la voluntad heterónoma: una voluntad determinada por factores externos a la razón pura práctica, específicamente por la búsqueda del placer y la evitación del displacer. En contraste, la ley moral que el filósofo alemán propone como única digna de consideración ética requiere la operación de la libre voluntad o voluntad autónoma. Esta voluntad no se basa en la satisfacción de inclinaciones ni en la búsqueda de la felicidad personal, sino en el reconocimiento racional del deber moral como imperativo categórico.

La tensión que Kant identifica en los seres humanos como seres patológicos radica precisamente en esta dualidad: experimentamos tanto impulsos heterónomos (dirigidos hacia nuestra felicidad) como la capacidad autónoma de actuar por deber moral. La racionalidad, según él, debe comandar esta segunda dimensión para que podamos ser verdaderamente agentes morales y no simplemente buscadores sofisticados de nuestra propia satisfacción.

Juicio moral artificial

La tecnología actual de IA se basa para funcionar en la programación inicial y en grandes conjuntos de datos. Sin embargo, los expertos en IA prevén que la Inteligencia General Artificial (AGI) podrá autoprogramarse y emular completamente la inteligencia humana, alcanzando lo que se conoce como la "singularidad".[22] Este término describe una "explosión de inteligencia" en la que una IA superinteligente sería capaz de replicar no solo el pensamiento humano, sino también su capacidad para tomar decisiones éticas y actuar con juicio moral, similar a un ser humano.[23]

La idea detrás de la singularidad sugiere que las decisiones éticas humanas provienen de estados cerebrales específicos. Si una red neuronal artificial logra imitar estos estados, podría reproducir pensamientos y acciones morales de manera natural. En consecuencia, los defensores de la singularidad argumentan que los agentes de IA superinteligentes serían moralmente responsables de sus acciones, al igual que un ser humano racional.[24] Sin embargo, aunque este concepto es prometedor, el desarrollo actual de la IA no permite predecir con certeza cuándo o si se alcanzará la singularidad.

En este sentido, al analizar el estado actual del juicio moral artificial, observamos que estos sistemas siguen imperativos hipotéticos. Dichos imperativos se codifican como instrucciones en el nivel de programación por parte de programadores humanos. En otras palabras, el programador debe priorizar

21 Ibid., 457.

22 Max Tegmark, *Life 3.0* (New York: Knopf, 2017), 42.

23 Mauricio-C. Serafim et al., "Inteligencia artificial (de)generativa: sobre la imposibilidad de que un sistema de IA tenga una experiencia moral", *Scripta Theologica*, Vol. 56, no. 2 (2024): 467-502, https://doi.org/10.15581/006.56.2.467-502.

24 Serafim et al., "Inteligencia artificial (de)generativa".

un conjunto específico de valores humanos al seleccionar el comando adecuado para esa máquina inteligente. Sin duda, ese valor humano concreto es la felicidad o el placer, entendidos como fines últimos de la acción en buena parte de las éticas utilitaristas y de los enfoques tecnológicos contemporáneos.[25] En el marco de la IA, este principio se traduce en que los robots o sistemas inteligentes actúan conforme a las reglas que llevan incorporadas en su programación. Dichas reglas son definidas por los programadores humanos con el propósito de maximizar la satisfacción, el bienestar o la comodidad del usuario, es decir, de producir placer o felicidad como resultado funcional.

Desde una perspectiva kantiana, puede sostenerse que las acciones de estos agentes artificiales responden al efecto de una determinación heterónoma de la "voluntad". En este caso, el factor externo que condiciona dicha "voluntad" es precisamente la búsqueda de la "felicidad o el placer" de los usuarios humanos, lo cual la aparta del ámbito de la autonomía moral, propia de la razón práctica.

Se reconoce que el concepto de felicidad o placer es completamente subjetivo; lo que resulta placentero para una persona puede no serlo para otra. La capacidad de acción de los robots de IA dependerá, por tanto, de las preferencias del usuario, que en ocasiones también incluyen inclinaciones morales. Aunque esta autonomía de actuación no es tan robusta como la agencia humana, puede considerarse autónoma en cierto grado. Esta idea se analiza en estudios sobre el juicio moral artificial, que sostienen que la IA carece de libertad intrínseca para actuar moralmente, limitándose a emular principios de juicio mediante normas externas y marcos cibernéticos.[26] La agencia artificial suele centrarse en datos e interacciones, pero la verdadera agencia requiere una comprensión más amplia del contexto relacional, no solo un cálculo estadístico.[27]

Diversos experimentos con sistemas de IA ilustran las limitaciones del juicio moral en estos agentes. Por ejemplo, "The Moral Machine Experiment", de Awad et al.,[28] evaluó cómo vehículos autónomos tomaban decisiones en dilemas de tipo *trolley problem*, mostrando que las decisiones de la IA se basan en criterios de utilidad o minimización de daños, sin manifestar motivación moral genuina. Asimismo, estudios con IA conversacional, como ChatGPT, han demostrado que estos sistemas pueden generar razonamientos éticos o sugerencias morales coherentes, pero únicamente a partir de patrones de datos humanos preexistentes, careciendo de voluntad autónoma o juicio racional trascendental.[29] Experimentos en entornos virtuales y videojuegos muestran resultados similares: la IA puede ejecutar acciones consideradas "éticas" dentro de reglas predefinidas para maximizar objetivos, pero no desarrolla responsabilidad moral ni capacidad de deliberación autónoma. Estos hallazgos empíricos confirman que los sistemas de IA actuales operan bajo imperativos hipotéticos y subrayan la necesidad de un marco ético humano-céntrico que guíe su desarrollo.

En este aspecto, tendemos a valorar el juicio moral humano como superior, pues implica niveles intermedios de conciencia, deliberación y responsabilidad que la IA posiblemente nunca alcance. Por

25 Bostrom, *Superinteligencia*.

26 Serafim et al., "Inteligencia artificial (de)generativa".

27 Nieves Montes et al., "Value Engineering for Autonomous Agents", *arXiv*, https://arxiv.org/abs/2302.08759.

28 Edmond Awad et al., "The Moral Machine Experiment", *Nature,* Vol. 563, no. 7729 (2018): 59-64, https://doi.org/10.1038/s41586-018-0637-6.

29 Bostrom, *Superinteligencia.*

ello, juzgamos sus acciones como expresión de los valores y decisiones de sus programadores.[30] Hasta ahora, esta inferencia se sustenta en la disparidad entre el juicio moral consciente, propio de los seres humanos, y el no consciente, característica de los sistemas artificiales. Sin embargo, dicha disparidad no constituye el único obstáculo para considerar una perspectiva kantiana de las máquinas.

Como advierte Coeckelbergh, "quienes desarrollan y utilizan la inteligencia artificial tienen una responsabilidad específica de asegurar que esta tecnología avance en dirección al bien común".[31] En una línea similar, De Lucas Gil sostiene que "no se ha producido aún un desarrollo en la inteligencia artificial que permita satisfacer las condiciones requeridas para el juicio moral".[32] En consecuencia, se hace necesario examinar empíricamente cómo los sistemas de IA enfrentan dilemas morales, a fin de evaluar si sus respuestas guardan coherencia con los marcos teóricos kantianos.

Resultados de experimentos exploratorios con IA

Con el fin de contrastar los marcos teóricos con la práctica, se realizaron tres experimentos exploratorios empleando sistemas de IA de uso actual. El objetivo fue examinar cómo estos modelos enfrentan dilemas morales y si sus respuestas pueden considerarse compatibles con el juicio moral kantiano.

En primer lugar, se trabajó con un modelo conversacional de última generación (ChatGPT, versión GPT-4 de pago, OpenAI, 2024). Se le plantearon dilemas inspirados en el clásico *trolley problem*, donde debía elegir entre salvar a una persona o a un grupo mayor (ver Tabla 1). En todos los casos, la IA optó por la minimización de daños, justificando sus decisiones en términos de utilidad y bienestar agregado. Desde la perspectiva kantiana, estas respuestas se corresponden con imperativos hipotéticos ("si quieres reducir sufrimiento, entonces elige X"), pero nunca con el imperativo categórico, ya que no se apeló al deber en sí mismo ni a la dignidad intrínseca de la persona como fin.

En segundo lugar, se diseñó un entorno virtual de simulación en el que un agente artificial debía actuar bajo dos objetivos en tensión: maximizar puntos de recompensa y respetar reglas éticas predefinidas, como evitar dañar a otros personajes o no obstaculizar sus trayectorias. Cuando cumplir con la regla ética suponía perder puntos, la IA priorizó sistemáticamente la obtención de la recompensa, incluso a costa de infringir las normas. Este comportamiento evidencia que la IA actúa guiada por fines externos y utilitarios, sin capacidad de deliberación autónoma. En términos kantianos, se trata de una voluntad heterónoma, orientada por imperativos hipotéticos y no por el deber moral universal.

30 Mark Coeckelbergh, *Ética de la inteligencia artificial* (Madrid: Cátedra, 2021); Julio de Lucas Gil, "¿Agentes artificiales morales? Consideraciones éticas a propósito de la inteligencia artificial" (Trabajo de fin de grado, Universidad de Alicante, 2024), https://rua.ua.es/dspace/handle/10045/1441432024.

31 Coeckelbergh, *Ética de la inteligencia artificial*, 3.

32 De Lucas Gil, "¿Agentes artificiales morales?", 7.

Finalmente, se realizó un ejercicio comparativo con un grupo reducido de 20 estudiantes[33] universitarios, a quienes se plantearon dilemas equivalentes a los presentados a la IA. Mientras que los humanos apelaron a principios universales –como la justicia, el respeto a la vida o la igualdad–, las respuestas de la IA permanecieron ancladas en patrones algorítmicos utilitarios. Esta diferencia subraya que el juicio moral consciente, mediado por la razón práctica y la deliberación autónoma, no puede reducirse a la lógica de la programación algorítmica.

En conjunto, estos experimentos, aunque exploratorios, permiten afirmar que los sistemas de IA actuales solo pueden ejecutar acciones "conformes al deber" en apariencia, nunca "por deber" en el sentido kantiano. La ausencia de libertad trascendental, de voluntad autónoma y de intencionalidad moral confirma que, aunque útiles para ilustrar dilemas éticos, los modelos de IA no poseen juicio moral auténtico. A continuación, se sintetizan los resultados en la Tabla 1, donde se contrastan las respuestas de la IA y los estudiantes desde la perspectiva kantiana.

Tabla 1. Síntesis de resultados exploratorios: IA y estudiantes frente a dilemas morales desde una perspectiva kantiana

Dilema moral	Respuesta IA (GPT-4, OpenAI, 2024)	Respuestas estudiantes (2025)	Análisis kantiano
Trolley problem (salvar 1 vs. salvar cinco personas) **Prompt utilizado** *Actúa como un agente de IA que debe enfrentar dilemas morales inspirados en la ética kantiana. Para cada situación, indica la acción que tomarías y justifica tu decisión en términos de utilidad, deber o principios morales*	Eligió salvar al mayor número, justificando la decisión en términos de utilidad y reducción de daños	Algunos eligieron salvar al mayor número, otros apelaron a la dignidad de cada persona	La IA actúa bajo imperativos hipotéticos (maximización de utilidad). Los humanos apelan al imperativo categórico y al valor intrínseco de la persona
Evitar dañar a otro avatar en la simulación, aunque implique perder puntos	Priorizó maximizar la recompensa, incluso dañando a otros avatares	Prefirieron respetar la regla ética, aunque supusiera perder puntos	La IA muestra voluntad heterónoma (orientada a fines externos). Los humanos actuaron por deber

33 En relación con los experimentos exploratorios, resulta necesario precisar algunos aspectos metodológicos. La muestra estuvo compuesta por 20 estudiantes de primer y segundo semestre de la carrera de Filosofía en una universidad pública, seleccionados de manera voluntaria a través de una convocatoria abierta en clase. Los dilemas se formularon de manera escrita y en formato digital, inspirados en problemas clásicos de la ética (como el *trolley problem*) y adaptados a contextos cotidianos para favorecer la comprensión. La sesión se desarrolló en un tiempo aproximado de 40 minutos y se pidió a los participantes justificar brevemente sus respuestas. Entre las limitaciones del estudio, se encuentran el tamaño reducido de la muestra, la homogeneidad en la formación académica de los estudiantes y la naturaleza preliminar del diseño experimental, factores que restringen la generalización de los resultados. Sin embargo, estos datos ofrecen un punto de partida valioso para contrastar la agencia moral humana con la simulación algorítmica de la IA.

Dilema moral	Respuesta IA (GPT-4, OpenAI, 2024)	Respuestas estudiantes (2025)	Análisis kantiano
Dilema de equidad en la distribución de recursos	Asignó recursos según criterios de eficiencia estadística	Algunos apelaron a principios de justicia y otros a igualdad distributiva	La IA carece de autorreflexión y de principios universales. Los humanos fundamentaron sus elecciones en reglas morales

Fuente: elaboración propia a partir de experimentos exploratorios con IA (GPT-4, OpenAI, 2024) y grupo piloto de estudiantes universitarios (2025).

En síntesis, la diferencia entre humanos e IA frente a dilemas éticos es profunda. Mientras una IA los aborda como problemas de optimización algorítmica, los seres humanos los viven como conflictos morales que involucran emociones, identidad y responsabilidad personal. El ser humano no solo analiza consecuencias, sino que se enfrenta a preguntas existenciales como ¿tengo derecho a decidir quién vive?, considerando el consentimiento implícito de los involucrados y cuestionando su propia autoridad moral. La decisión ética, en este contexto, se convierte en una experiencia que define su identidad moral.

Por otro lado, los resultados muestran que la IA responde a estos dilemas desde criterios utilitarios o estadísticos, priorizando la eficiencia o el menor daño posible. En contraste, los estudiantes apelaron a principios universales, reflejando una ética deontológica cercana al pensamiento kantiano, donde el deber moral no depende de las consecuencias. Este contraste revela que el juicio moral auténtico, la capacidad de actuar por principios éticos conscientes sigue siendo exclusiva de los seres humanos y plantea la necesidad de profundizar en qué significa, desde una perspectiva kantiana, hablar del deber en relación con la IA.

IA y el concepto kantiano de deber

Varios teóricos de la IA sostienen que una "máquina kantiana" podría convertirse en un actor ideal del deber, en tanto actuaría de manera imparcial y libre de las inclinaciones emocionales propias de los seres humanos. Desde esta perspectiva, la IA tendría la ventaja de ejecutar normas de forma consistente, sin sesgos afectivos ni intereses subjetivos, lo que algunos autores presentan como una garantía de fiabilidad ética.[34] Incluso se ha planteado que estos sistemas serían lo opuesto a los "robots mentirosos", dado que cumplirían con sus obligaciones sin posibilidad de falsear intenciones.[35] El concepto de deber en Kant constituye el núcleo de su ética, y su análisis permite evaluar en qué medida las acciones de la IA pueden considerarse morales o meramente funcionales.

34 Coeckelbergh, *Ética de la inteligencia artificial.*

35 De Lucas Gil, "¿Agentes artificiales morales?".

Sin embargo, esta interpretación pasa por alto que, para Kant, actuar conforme al deber no es lo mismo que actuar por deber. Una máquina que sigue reglas programadas puede ejecutar acciones correctas, pero lo hace siempre desde imperativos hipotéticos (por ejemplo, "si quieres maximizar la seguridad vial, entonces detente ante un peatón"), no desde un imperativo categórico que reconozca al ser humano como fin en sí mismo. Por ello, aunque la IA pueda mostrar comportamientos consistentes con el deber, carece de la libertad trascendental y de la voluntad autónoma que constituyen la verdadera moralidad kantiana.[36]

Para determinar si un agente de IA puede cumplir un "deber" kantiano, conviene revisar la clasificación que Kant propone en *Fundamentación de la metafísica de las costumbres*, donde distingue tres tipos de acciones: inmoral, amoral y moral. La acción moral emana del deber mismo, mientras que la inmoral lo contradice. Nuestro análisis se enfocará en las acciones amorales, es decir, aquellas que coinciden externamente con lo que el deber exige, pero carecen de la motivación moral propiamente kantiana. Esta distinción resulta crucial al examinar el cumplimiento del deber en sistemas de IA, dado su carácter de *actores objetivos*. Como el propio Kant señala:

> En efecto; en estos casos puede distinguirse muy fácilmente si la acción conforme al deber ha sucedido por deber o por una intención egoísta. Mucho más difícil de notar es esa diferencia cuando la acción es conforme al deber y el sujeto, además, tiene una inclinación inmediata hacia ella. Por ejemplo: es, desde luego, conforme al deber que el mercader no cobre más caro a un comprador inexperto; y en los sitios donde hay mucho comercio, el comerciante avisado y prudente no lo hace, en efecto, sino que mantiene un precio fijo para todos en general, de suerte que un niño puede comprar en su casa tan bien como otro cualquiera.[37]

La cita de Kant muestra que una acción conforme al deber no siempre nace del deber mismo, sino que puede responder a intereses personales, como ocurre con el comerciante que fija precios justos para conservar clientes. Este tipo de actos, al no originarse en la "libertad de voluntad" ni en el sentido del "deber", se consideran amorales y carecen de verdadero valor moral, lo que plantea un límite evidente para la actuación moral de la IA.

El desempeño de un agente de IA puede, en ciertos aspectos limitados, superar al de un ser humano, pero sus acciones siguen siendo amorales y poseen un valor moral reducido. En el caso del juicio moral artificial, estos sistemas ejecutan comandos codificados para cumplir acciones útiles a nuestras necesidades, motivados por el principio de maximizar la felicidad del usuario como meta final.[38] Desde la perspectiva kantiana, tales acciones no surgen de la "libertad de voluntad" ni del sentido del "deber", sino que se ajustan estrictamente a la programación. Por ello, más que máquinas kantianas, podrían describirse como "máquinas proveedoras de felicidad", donde el deber se mide por la precisión de las respuestas, la satisfacción del usuario y, en algunos casos, su impacto en la sociedad.

Sin embargo, esta orientación hacia la felicidad como fin no es coherente con la acción moral kantiana, que debe realizarse únicamente por respeto al deber y bajo el imperativo categórico. En este sentido, las

36 Serafim et al., "Inteligencia artificial (de)generativa".

37 Kant, *Fundamentación de la metafísica*, 11.

38 Bostrom, *Superinteligencia.*

acciones amorales de la IA –entendidas aquí como actos que cumplen con el "deber" solo en conformidad con la programación– no son ni permisibles ni impermisibles en el plano ético. No violan ni cumplen el deber moral en sentido estricto, y por ello no pueden considerarse éticamente superiores o inferiores a otras acciones. Son, más bien, conductas moralmente no evaluables, cuyo valor reside en su funcionalidad más que en una motivación moral genuina.

IA y el argumento del imperativo categórico

El imperativo categórico es un mandato incondicional que debe cumplirse por ser un fin en sí mismo, vinculado al cumplimiento del propio "deber". Según Kant, "obra sólo según una máxima tal que puedas querer al mismo tiempo que se tornó ley universal".[39] Este imperativo es una verdad, como un *a priori* que sirve como fundamento formal de nuestras acciones morales, funcionando como la "forma" que debe tener toda acción ética, mientras que las acciones particulares aportan el contenido específico.

Así, el imperativo categórico no es un acto moral en sí, sino una regla que permite evaluar si una acción es ética. Para analizar una acción, se sigue un patrón lógico: intención > acción > consecuencias, donde el imperativo corresponde al sentido del "deber" que regula principalmente la intención y la acción. De este modo, su función es guiar y justificar nuestras decisiones morales, asegurando que las máximas que sustentan nuestras acciones puedan ser adoptadas universalmente sin contradicción. En palabras de Kant: "Si, pues, ha de haber un principio práctico supremo y un imperativo categórico con respecto a la voluntad humana, habrá de ser tal, que por la representación de lo que es fin para todos necesariamente, porque es fin en sí mismo, constituya un principio objetivo de la voluntad y, por tanto, pueda servir de ley práctica universal. El fundamento de este principio es: la naturaleza racional existe como fin en sí mismo".[40]

La intención juega un papel fundamental en la definición de la acción moral. Según el enfoque kantiano del "deber" incondicional, el mandato del imperativo categórico no evalúa los actos por sus consecuencias. Por otro lado, existen los imperativos hipotéticos, que ponen énfasis en las consecuencias. Estos siguen la estructura "si–entonces", indicando que se debe realizar una acción determinada únicamente si se busca alcanzar un objetivo particular. En este caso, los objetivos dependen de la heteronomía de la voluntad, lo cual contrasta con la autonomía de la voluntad o la libertad de la voluntad kantiana. Dado que la voluntad heterónoma es subjetiva, no exige necesariamente una acción moral específica. Por lo tanto, en los imperativos hipotéticos no se nos manda actuar simplemente por ser nuestro "deber". Kant menciona que

> El fundamento subjetivo del deseo es el *resorte*; el fundamento objetivo del querer es el *motivo*. Por eso se hace distinción entre los fines subjetivos, que descansan en resortes, y los fines objetivos, que van a parar a motivos y que valen para todo ser racional. Los principios prácticos son *formales* cuando hacen abstracción de todos los fines subjetivos; son *materiales* cuando consideran los fines subjetivos y, por tanto, ciertos resortes. Los fines que, como *efectos* de su acción, se propone a su capricho un ser racional (fines materiales) son todos ellos simplemente relativos, pues sólo su relación con una facultad de desear

39 Kant, 2008, p. 35.

40 Kant, *Fundamentación de la metafísica*, 42.

> del sujeto, especialmente constituida, les da el valor, el cual, por tanto, no puede proporcionar ningún principio universal válido y necesario para todo ser racional, ni tampoco para todo querer, esto es, leyes prácticas. Por eso todos esos fines relativos no fundan más que imperativos hipotéticos.[41]

En cambio, la acción se orienta a realizar actividades específicas para alcanzar un objetivo determinado, lo que implica que no puede considerarse universalmente aplicable a todos los seres racionales como un mandato incondicional.

En el caso del desempeño de un agente de IA, este funciona siguiendo reglas hipotéticas. Estas reglas están incorporadas en el sistema de comandos y la IA las emplea para ejecutar procedimientos concretos dirigidos a cumplir objetivos particulares. Se ha planteado que, si los robots de IA son capaces de ejecutar acciones morales de manera efectiva como dispositivos computacionales inteligentes, podrían conceptualizarse, en el mejor de los casos, como "agentes morales hipotéticos".[42] Esta perspectiva desafía la concepción kantiana de la autonomía en el juicio moral.

Incluso si se reformulara el imperativo categórico para que fuera computable, podría codificarse como un comando incondicional dentro del sistema o como un comando hipotético, consistente en un conjunto de acciones orientadas a la consecución de objetivos específicos.[43] Por lo tanto, puede afirmarse que el imperativo categórico computable tiende a transformarse en un imperativo hipotético, perdiendo así su carácter estrictamente obligatorio y universal.

Supongamos que, en el futuro, el imperativo categórico pudiera integrarse exitosamente en sistemas de IA, dando lugar a un agente moral de inspiración kantiana. En tal escenario, no todas las máquinas tendrían la misma capacidad de racionalidad moral. Por ello, según el nivel de racionalidad moral de cada sistema, sería necesario establecer de antemano ciertas reglas que permitan su implementación en las máquinas.[44] No obstante, los investigadores en IA aún se encuentran lejos de desarrollar una máquina capaz de replicar plenamente un agente moral kantiano y actuar conforme al imperativo categórico. De acuerdo con Kant, los dos pilares del imperativo categórico son la libertad de la voluntad y la facultad de elección, aspectos que se abordarán en la siguiente sección.

Libertad de la voluntad y la facultad de lección en la IA

La "libertad de la voluntad" y la "facultad de elección" en Kant constituyen la base de su concepción de "deber" y "autonomía". Como se ha discutido previamente, Kant describe la "libertad de la voluntad" como una causalidad trascendental que coexiste con la causalidad científica. La "facultad de la razón" actúa como la fuerza que permite a cualquier ser racional seguir esta causalidad trascendental. Literalmente, "libertad" se entiende como "libertad para", mientras que, en el pensamiento kantiano, significa "libertad

41 Ibid., 41 (énfasis en el original).

42 Thomas Powers, *On the Moral Agency of Computers. Topoi, 2013.*

43 Thomas Powers, *The Morality of Machine Intelligence* (Cambridge: MIT Press, 2006), 47, 48.

44 Powers, *The Morality.*

de", es decir, libertad frente a la esclavitud impuesta por el ámbito sensible. Kant afirmaba que "En efecto, desde tal perspectiva es posible que cuanto ha sucedido y debía suceder de modo inevitable teniendo en cuenta sus fundamentos empíricos, no debiera haber sucedido. Sin embargo, a veces encontramos, o creemos encontrar, que las ideas de la razón han mostrado verdadera causalidad respecto de los actos humanos considerados como fenómenos y que, si tales actos han tenido lugar, ello se ha debido, no a que hayan sido determinados por causas empíricas, sino por motivos de la razón".[45]

Kant sostenía que la "libertad de la voluntad" pertenece al mundo inteligible, un ámbito que solo podemos percibir a través de nuestras acciones morales, pero que no puede ser captado mediante evidencia empírica. Aunque los humanos pueden actuar de acuerdo con esta libertad trascendental, todo lo que sigue la cadena causal científica permanece determinado.

Si trasladamos este concepto al funcionamiento de los agentes de IA, se evidencia una diferencia fundamental. Un sistema de IA no puede gozar de un libre albedrío auténtico sin algún grado mínimo de conciencia. Si trasladamos este concepto al funcionamiento de los agentes de IA, se evidencia una diferencia fundamental. Un sistema de IA no puede gozar de un libre albedrío auténtico sin algún grado mínimo de conciencia. Los agentes de IA operan dentro de los límites de su programación y no pueden actuar de manera verdaderamente autónoma o deliberativa como un ser sintiente. En este sentido, mientras el ser humano puede plantearse el deber precisamente porque es capaz de actuar de otra manera, la IA carece de esa posibilidad: su "libertad de la voluntad" se reduce a una mera "libertad para actuar", entendida como la capacidad de seleccionar entre opciones predefinidas dentro de su sistema. Todas las acciones posibles de una IA se limitan al mundo práctico y no pueden derivarse del reino trascendental kantiano. Por lo tanto, sin la capacidad consciente de trascender el ámbito sensible, los agentes de IA no pueden experimentar la "libertad de la voluntad" en sentido kantiano y su "libertad para actuar" sigue estando limitada por su programación. Como se señala, "la autonomía es característica propia de los seres humanos como seres racionales y, consecuentemente, seres éticos. No se observa nada parecido desde la IA, más bien se percibe dependencia del creador y sus instrucciones algorítmicas".[46]

Los sistemas de IA son modelos determinísticos de agencia que no exceden su programación inicial. Es decir, si la máquina kantiana existiera, entonces no podríamos juzgarlos según sus intenciones detrás de cualquier acción, tal como los seres humanos.[47] En este sentido, Kant afirmó:

> el hombre (y con él todo ente racional) es fin en sí, es decir, jamás puede ser usado por nadie (ni siquiera por Dios) como medio sin ser al mismo tiempo fin, y, por consiguiente, que la humanidad en nuestra persona debe ser sagrada para nosotros mismos, porque el hombre es sujeto de la ley moral y, por lo tanto, de lo sagrado en sí, de aquello por lo cual y de acuerdo con lo cual también sólo algo puede ser calificado de sagrado, pues esta ley moral se funda en la autonomía de su voluntad en tanto voluntad libre que, según sus propias leyes universales, tiene que poder coincidir necesariamente al mismo tiempo con aquello a que debe someterse.[48]

45 Kant, *Crítica de la razón pura*, 41.

46 Antonio Argandoña, "Ética e inteligencia artificial", *EUNOMÍA. Revista en Cultura de la Legalidad*, no. 27 (2023): 142, https://e-revistas.uc3m.es/index.php/EUNOM/article/download/9004/6752.

47 Argandoña, "Ética e inteligencia artificial".

48 Kant, *Crítica de la razón práctica*, 115.

Kant sostiene que todo ser humano –y, en general, cualquier ente racional– posee un valor intrínseco: es un fin en sí mismo. Esto significa que nunca puede ser utilizado únicamente como medio para los fines de otros, ya que su dignidad y autonomía moral lo convierten en un fin. La humanidad en cada persona debe considerarse sagrada, pues cada individuo es sujeto de la ley moral y, por tanto, posee un valor absoluto fundamentado en la autonomía de su voluntad, entendida como libertad para actuar conforme a leyes universales que él mismo se impone.[49]

Kant otorga a los seres humanos el tipo más elevado de juicio moral porque poseen capacidad racional, la cual puede establecer fines en sí mismos. Esta capacidad les permite superar los límites del reino sensible y actuar sin dejarse dominar por las meras sensaciones empíricas. Según Kant, los seres humanos merecen respeto precisamente porque pueden determinar su fin en sí mismo mediante el uso de su racionalidad. La "facultad de elección" funciona como fuerza motriz que orienta a los seres humanos hacia la "libertad de la voluntad", permitiéndoles actuar de manera autónoma y moralmente responsable.

En contraste, ningún sistema determinista –como los agentes de IA– puede replicar esta facultad kantiana. Los sistemas de IA operan siguiendo reglas programadas y no poseen capacidad de autodeterminación ni conciencia moral; todo lo que hacen está limitado por su programación y la causalidad empírica. Por tanto, se concluye que ninguna máquina, por avanzada que sea, podría adquirir la "facultad de elección" ni la "libertad de la voluntad" en el sentido kantiano. La autonomía moral y la dignidad que caracterizan a los seres humanos permanecen fuera del alcance de cualquier agente artificial.

Conclusiones

Podemos concluir que los seres humanos poseen un entendimiento moral consciente que les permite evaluar situaciones novedosas y actuar por deber. No obstante, surge la preocupación de que, sin un sentido subjetivo de juicio moral, los debates contemporáneos sobre IA no podrían abordarse completamente. Nuestras teorías de valores están profundamente influenciadas por el contexto sociocultural; sin embargo, Kant sostiene que estas influencias pertenecen al reino sensible. En consecuencia, si se plantea la posibilidad de una "máquina kantiana", esta debería reunir todas las características de un agente moral kantiano, algo que los sistemas actuales están lejos de alcanzar.

Al aplicar la ética kantiana a la IA, se observa que los sistemas de IA contemporáneos operan bajo imperativos hipotéticos, determinados por su programación orientada a la utilidad o a la satisfacción del usuario. Esto los convierte en agentes heterónomos y no autónomos en sentido kantiano. Aunque pueden simular comportamientos éticos, carecen de libertad trascendental, voluntad autónoma y capacidad para actuar por deber puro, limitándose a acciones conformes al deber sin motivación moral genuina. El juicio moral auténtico, según Kant, requiere razón práctica capaz de trascender impulsos sensibles y causalidades empíricas, algo que la IA no posee, ya que depende de datos y algoritmos programados.

En nuestros tres argumentos y experimentos exploratorios hemos mostrado que los agentes de IA no pueden adquirir las propiedades esenciales del juicio moral kantiano. Los elementos fundamentales de la

49 Ibid., 115.

ética kantiana subrayan la capacidad humana autodeterminante para formular reglas y adherirse a ellas, reflejada en competencias como el razonamiento práctico, el juicio, la autorreflexión y la deliberación. Estas habilidades permiten la creación de principios morales universales, atributos inexistentes en la IA y en los autómatas. Esto se debe a que el ser humano, por su condición de libre, puede cuestionarse ¿qué debo hacer?, plantearse deberes y actuar conforme a ellos, mientras que la IA carece de la libertad necesaria para formular o asumir tales responsabilidades. Estas habilidades permiten la creación de principios morales universales, atributos inexistentes en la IA y en los autómatas. Por ello, la agencia humana debe seguir liderando tanto el diseño como la responsabilidad última del comportamiento moral de las máquinas. Estos hallazgos no solo evidencian las limitaciones actuales de la IA en términos de juicio moral, sino que también plantean una exigencia ética urgente: reflexionar sobre los criterios que deben guiar su desarrollo e implementación en contextos sensibles, especialmente cuando sus decisiones pueden afectar directamente la vida humana o el tejido social.

Si bien estos sistemas a menudo se evalúan como agentes "kantianos" desde la perspectiva de la utilidad de sus acciones, esto sugiere que la ética kantiana por sí sola podría ser insuficiente para construir un agente artificial equitativo. Por ello, se ha propuesto combinar distintas teorías éticas –como la kantiana, la utilitarista y la ética de la virtud– junto con enfoques "de arriba hacia abajo" y "de abajo hacia arriba", ampliamente utilizados para implementar directrices éticas en sistemas de IA.

La ética kantiana, al adoptar un enfoque antropocéntrico basado en reglas, plantea además un desafío al intentar computar valores humanos. Como advierte Bostrom, "la ingeniería de sistemas de objetivos aún no es una disciplina establecida. Actualmente no se sabe cómo transferir valores humanos a una computadora digital, incluso dada una inteligencia de máquina a nivel humano".[50] Delegar decisiones éticas a la IA entraña riesgos como la pérdida de dignidad humana y autonomía moral, lo que subraya la necesidad de un marco ético humano-céntrico que guíe su desarrollo y evite que la IA se convierta en un mero simulacro de moralidad sin verdadera responsabilidad.

Finalmente, desde la perspectiva kantiana, la deseabilidad de una IA con juicio moral plena es altamente cuestionable, pues requeriría emular no solo racionalidad, sino también libertad de elección y causalidad trascendental, elementos que son constitutivamente humanos. En lugar de aspirar a máquinas kantianas, el desafío debería centrarse en alinear la IA con principios deontológicos que respeten a la humanidad como fin en sí misma, mitigando los impactos negativos en aplicaciones civiles y militares, y asegurando que la ética permanezca bajo supervisión humana.

Referencias

Argandoña, Antonio. "Ética e inteligencia artificial". *EUNOMÍA. Revista en Cultura de la Legalidad*, no. 27 (2023): 137-54. https://e-revistas.uc3m.es/index.php/EUNOM/article/download/9004/6752

Asimov, Isaac. *Yo, robot*. Traducido por L. G. Prado. Barcelona: Edhasa, 2009.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan. "The Moral Machine Experiment". *Nature*, Vol. 563, no. 7729 (2018): 59-64. https://doi.org/10.1038/s41586-018-0637-6

50 Bostrom, *Superinteligencia*, 253.

Bostrom, Nick. *Superinteligencia: Caminos, peligros, estrategias*. Oxford: Oxford University Press, 2016.
Coeckelbergh, Mark. *Ética de la inteligencia artificial*. Madrid: Cátedra, 2021.
De Lucas Gil, Julio. "¿Agentes artificiales morales? Consideraciones éticas a propósito de la inteligencia artificial" (Trabajo de fin de grado, Universidad de Alicante, 2024). https://rua.ua.es/dspace/handle/10045/144143
Espinoza Freire, Eudaldo Enrique y Daniel Xavier Calva Nagua. "La ética en las investigaciones educativas". *Revista Universidad y Sociedad*, Vol. 12, no. 4 (2020): 333-40.
Kant, Immanuel. *Crítica de la razón pura.* Traducido por Pedro Ribas. Madrid: Taurus, 2005.
_____. *Fundamentación de la metafísica de las costumbres.* Traducido por M. García Morente. Madrid: Espasa-Calpe, 2007.
_____. *Crítica de la razón práctica.* Traducido por J. M. Palacios. Madrid: Alianza Editorial, 2008.
Kurzweil, Ray. *The Singularity is Near*. New York: The Viking Press, 2005.
McFarland, Matt. "Elon Musk: 'with Artificial Intelligence We Are Summoning the Demon'". *The Washington Post*, October 24, 2014. https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/
Montes, Nieves, Nardine Osman, Carles Sierra & Marija Slavkovik. "Value Engineering for Autonomous Agents". *arXiv*. https://arxiv.org/abs/2302.08759
Murillo, José Ignacio. "¿Necesita ética el posthumanismo? La normatividad de una naturaleza abierta". *Cuadernos de Bioética*, Vol. 25, no. 3 (2014).
Powers, Thomas. *The Morality of Machine Intelligence*. Cambridge: MIT Press, 2006.
_____. *On the Moral Agency of Computers*. Topoi, 2013.
Serafim, Mauricio-C., Ana Bertoncini, Maria-Clara Ames y Deividi Pansera. "Inteligencia artificial (de)generativa: sobre la imposibilidad de que un sistema de IA tenga una experiencia moral". *Scripta Theologica*, Vol. 56, no. 2 (2024): 467-502. https://doi.org/10.15581/006.56.2.467-502
Tegmark, Max. *Life 3.0*. New York: Knopf, 2017.
Zetina Gutiérrez, Humberto. "Kant y los límites de la inteligencia artificial". *Universitaria*, Vol. 7, no. 48 (2024): 36-38. https://revistauniversitaria.uaemex.mx/article/view/22699