Tipo de artículo: Artículo original

Comprendiendo los límites de la automatización moral

Understanding the limits of moral automation

Mario González Arencibia [1*], **https://orcid.org/0000-0001-9947-7762**
Omar Mar Cornelio [2], **https://orcid.org/0000-0002-0689-6341**

[1] Centro de Estudios de Gestión de Proyectos y Toma de Decisiones. Universidad de las Ciencias Informáticas. Cuba mgarencibia@uci.cu
[2] Centro de Estudio de Matemática Computacional. Universidad de las Ciencias Informáticas. Cuba. omarmar@uci.cu

*** Autor para correspondencia: mgarencibia@uci.cu**

Resumen

La automatización de decisiones morales es un tema que suscita profundo interés en la sociedad contemporánea, caracterizada por un creciente uso de la tecnología. Esta investigación se propone analizar los factores cruciales que deben considerarse al establecer límites para la automatización de juicios y acciones con implicaciones éticas. Los investigadores se plantearon la siguiente pregunta central: ¿Qué tipos de decisiones éticas no deberían delegarse a sistemas automatizados y cómo se puede mantener la responsabilidad humana en procesos morales automatizados? El estudio reveló que existen categorías específicas de decisiones éticas que no deberían ser completamente automatizadas. Entre estas se encuentran aquellas que implican dilemas morales complejos, las que afectan directamente la vida humana, o las que requieren un alto grado de empatía y comprensión contextual. Además, los hallazgos subrayan la importancia de mantener la supervisión y responsabilidad humana, incluso en procesos que están parcialmente automatizados. Los investigadores concluyeron que es fundamental alcanzar un equilibrio entre aprovechar la eficiencia que ofrece la automatización y preservar el juicio ético humano. Esto implica el diseño de sistemas capaces de identificar situaciones que requieren intervención humana y el establecimiento de mecanismos claros de rendición de cuentas. En síntesis, el estudio determina que, si bien la automatización presenta ventajas en términos de eficiencia y consistencia, existen límites éticos que deben ser respetados. Las decisiones morales más complejas y con mayor impacto en la vida humana deben permanecer bajo control humano, mientras que la automatización puede aplicarse de forma limitada en procesos más rutinarios, siempre bajo supervisión humana y con mecanismos de responsabilidad claramente definidos.

Palabras clave: Toma de decisiones, Principios éticos, Incertidumbre, Supervisión humana, Aprendizaje ético, Responsabilidad.

Abstract

The automation of moral decisions is a topic that arouses deep interest in contemporary society, characterized by an increasing use of technology. This research aims to analyze the crucial factors that must be considered when establishing limits for the automation of judgments and actions with ethical implications. The researchers posed the following central question: What types of ethical decisions should not be delegated to automated systems, and how can human responsibility be maintained in automated moral processes? The study revealed that there are specific categories of ethical decisions that should not be fully automated. These include those involving complex moral dilemmas, those directly affecting human life, or those requiring a high degree of empathy and contextual understanding. Furthermore, the findings underscore the importance of maintaining human oversight and responsibility, even in processes that are partially automated. The researchers concluded that it is essential to achieve a balance between leveraging the efficiency offered by automation and preserving human ethical judgment. This involves designing systems capable of identifying situations that require human intervention and establishing clear accountability mechanisms. In summary, the study determines that while automation presents advantages in terms of efficiency and consistency, there are ethical limits that must be respected. The most complex moral decisions with the greatest impact on human life should remain under human control, while automation can be applied in a limited way to more routine processes, always under human supervision and with clearly defined responsibility mechanisms.

Keywords: *Decision-making, Ethical principles, Uncertainty, Human oversight, Ethical learning, Accountability.*
Recibido: 08/06/2024
Aceptado: 22/08/2024
En línea: 01/09/2024

Introducción

La automatización moral se ha convertido en un tema de creciente interés en el campo de la ética y la tecnología. La automatización moral se refiere al uso de sistemas automatizados y algoritmos para tomar decisiones que tradicionalmente han estado bajo el dominio del juicio humano, especialmente en contextos éticos y morales (Rueda, 2023). Este fenómeno es particularmente relevante en la toma de decisiones en áreas como la justicia penal, la atención médica y la gestión de recursos, donde los algoritmos pueden influir en resultados cruciales para la vida y el bienestar de las personas.

A medida que los sistemas automatizados se vuelven más avanzados y omnipresentes en la sociedad, surge la necesidad de comprender y definir los límites éticos de su implementación. Investigaciones previas han abordado aspectos relacionados con este tema. Por ejemplo, Wallach y Allen (2008) discuten cómo la automatización puede aumentar la eficiencia y precisión en la toma de decisiones, pero también alerta sobre la pérdida de juicio humano crítico en procesos éticos. Otro trabajo de Binns et al. (2018) aborda las dificultades inherentes en la programación de algoritmos para tomar decisiones morales, subrayando la falta de flexibilidad y comprensión contextual que caracteriza a las máquinas en comparación con los humanos.

En este sentido, gobiernos e instituciones internacionales se han pronunciado sobre la necesidad de regular la automatización moral. La Comisión Europea, por ejemplo, ha propuesto directrices para garantizar que la inteligencia artificial se desarrolle y utilice de manera ética y transparente (European Commission, 2021). De manera similar, el Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) ha desarrollado principios éticos para el diseño y la implementación de sistemas autónomos, que enfatizan la importancia de la responsabilidad humana en el proceso de toma de decisiones (IEEE, 2020).

Organismos internacionales como la UNESCO han destacado la importancia de desarrollar marcos éticos para la inteligencia artificial y los sistemas automatizados. En su informe "Artificial Intelligence and Ethics" (2019), la UNESCO enfatiza la necesidad de salvaguardar los valores humanos en el diseño e implementación de tecnologías automatizadas.

Sin embargo, cuando se hace el balance del tema en la literatura internacional la mayoría de los estudios se han centrado más en los beneficios generales de la automatización, dejando de lado un análisis exhaustivo de las

implicaciones éticas y prácticas de su implementación en contextos diversos. La consecuencia es que existe una inadecuada comprensión de los límites de la automatización moral.

El objetivo principal de esta investigación es analizar los factores que deben considerarse al establecer límites para la automatización de decisiones morales. Se abordarán preguntas como: ¿Qué tipos de decisiones éticas no deberían delegarse a sistemas automatizados? ¿Cómo se puede mantener la responsabilidad humana en procesos morales automatizados?

Este estudio se justifica por la creciente adopción de sistemas automatizados en áreas sensibles como la atención médica, el sistema judicial y los vehículos autónomos, donde las decisiones pueden tener consecuencias significativas. Comprender los límites de la automatización moral es esencial para garantizar que estos sistemas operen de manera ética y responsable.

Materiales y métodos

El estudio de "Comprendiendo los Límites de la Automatización Moral" se basa en una revisión exhaustiva de la literatura académica, documentos normativos y estudios de caso recientes. La literatura revisada incluye trabajos clave sobre ética de la inteligencia artificial (IA) y algoritmos, como los de Floridi (2019) y O'Neil, (2016) que exploran la interacción entre la moralidad y la tecnología. Además, se analizaron documentos regulatorios, como el Acta de Inteligencia Artificial de la Comisión Europea y el Reglamento General de Protección de Datos, para evaluar cómo las políticas actuales abordan los problemas éticos de la IA.

El método principal consistió en un análisis crítico de los conceptos de automatización moral y ética algorítmica, complementado por el estudio de casos prácticos que ilustran los desafíos reales en la implementación de sistemas automatizados. Estos estudios de caso, como el de Amazon y el uso de algoritmos de policía predictiva, demostraron cómo la automatización puede replicar y amplificar sesgos humanos, resaltando las dificultades en la aplicación de principios éticos en contextos prácticos.

Las reflexiones críticas del análisis evidencian la insuficiencia de las regulaciones actuales para abordar completamente los desafíos éticos de la automatización moral. A pesar de las iniciativas regulatorias, como la legislación de la Unión Europea, persisten problemas significativos en la transparencia y la equidad de los sistemas de IA. Esto subraya la necesidad de enfoques más robustos para la ética algorítmica y una mayor transparencia en el diseño y uso de tecnologías automatizadas.

Resultados

Historia de la automatización en la toma de decisiones éticas

La historia de la automatización en la toma de decisiones éticas ha experimentado una evolución significativa, marcada por hitos que han configurado el panorama actual de esta disciplina (Consulte Tabla 1). En la década de 1950, con el surgimiento de la inteligencia artificial, comenzaron las primeras discusiones sobre la posibilidad de que las máquinas pudieran tomar decisiones éticas. El test de Turing, propuesto por Alan Turing en 1950, planteó la cuestión de si las máquinas podían exhibir un comportamiento inteligente indistinguible del de un ser humano, lo que sentó las bases para futuras consideraciones éticas en la automatización (Turing, 1950).

Tabla 1: Evolución del Concepto de Automatización Moral.

Etapa	Enfoque Principal	Características	Críticas	Fecha Clave
Automatización Industrial	Tareas repetitivas y predecibles.	- Eficiencia y precisión. - Reducción de errores humanos.	- Falta de consideración de aspectos éticos. - Deshumanización del trabajo.	1760s (Revolución Industrial)
Automatización en la Vida Cotidiana	Decisiones automatizadas con implicaciones morales.	- Conveniencia y personalización. - Potencial para mejorar la eficiencia y la justicia.	- Preocupaciones sobre la privacidad y la autonomía. - Riesgo de decisiones sesgadas o discriminatorias.	1970s-presente (Era de la Información)
Ascenso de la Inteligencia Artificial (IA)	Algoritmos para procesar datos y tomar decisiones.	- Potencial para mejorar la toma de decisiones. - Mayor objetividad y consistencia.	- Simplificación excesiva de la complejidad moral. - Subestimación del papel del juicio humano.	1950s-presente (Desarrollo de la IA)
Ética Algorítmica	Principios para el diseño y funcionamiento de algoritmos.	- Incorporación de consideraciones éticas en sistemas automatizados. - Promoción de la transparencia y la rendición de cuentas.	- Dificultad para replicar la flexibilidad y adaptabilidad de la ética humana.	2010s-presente (Creciente preocupación por la ética de la IA)
Interrelación entre Ética Algorítmica y Ética Humana	Búsqueda de un equilibrio entre ambos enfoques.	- Complementar la toma de decisiones humana con análisis de datos. - Asegurar que las decisiones automatizadas reflejen valores humanos.	Ética algorítmica no sustituye la ética humana.	2020s-presente (Enfoque en la colaboración entre la IA y la ética humana)
Automatización Moral Efectiva	Integración cuidadosa de la ética algorítmica y la ética humana.	- Diseño de algoritmos con comprensión profunda de los valores humanos. - Supervisión y ajuste humano constantes.	Automatización moral efectiva, no puede ser independiente a la práctica supervisión humana.	2020s-presente (Desarrollo de marcos éticos para la IA)

Fuente: Elaboración propia

Durante los años 60 y 70, el desarrollo de sistemas expertos en campos como la medicina y el derecho planteó nuevas preguntas sobre la responsabilidad y la toma de decisiones automatizadas. Estos sistemas, diseñados para emular el razonamiento de expertos humanos, pusieron de manifiesto la necesidad de considerar las implicaciones éticas de delegar decisiones importantes a máquinas.

Un hito significativo ocurrió en 1976 con la publicación del artículo "Computer Power and Human Reason" de Joseph Weizenbaum, que cuestionaba la idoneidad de utilizar computadoras para tomar decisiones en áreas que requieren comprensión y empatía humana. Este trabajo influyó en el debate sobre los límites éticos de la automatización (Weizenbaum, 1976).

En la década de 1980, el desarrollo de sistemas de soporte a la decisión más sofisticados llevó a una mayor conciencia sobre la necesidad de incorporar consideraciones éticas en el diseño de estos sistemas. La obra "The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World" de Edward Feigenbaum y Pamela McCorduck en 1983 estimuló el debate sobre las implicaciones éticas y sociales de la inteligencia artificial avanzada.

Los años 90 vieron un aumento en la investigación sobre ética computacional y el surgimiento de marcos teóricos para abordar los dilemas éticos en la automatización. El trabajo de James Moor, "What is Computer Ethics?" publicado en 1985, sentó las bases para esta disciplina emergente (Moor, 1985).

Con la llegada del siglo XXI, la rápida expansión de la inteligencia artificial y el aprendizaje automático ha intensificado el debate sobre la ética en la toma de decisiones automatizadas. En 2016, la Unión Europea aprobó el Reglamento General de Protección de Datos (GDPR), que incluye disposiciones sobre decisiones automatizadas y perfilado, marcando un hito en la regulación de la toma de decisiones automatizadas (European Union, 2016).

En 2018, el caso de Cambridge Analytica puso de manifiesto los riesgos éticos asociados con el uso de algoritmos para influir en la toma de decisiones políticas, lo que llevó a un mayor escrutinio público y regulatorio de las tecnologías de automatización (Cadwalladr, & Graham-Harrison, 2018).

Más recientemente, en 2020, la pandemia de COVID-19 ha acelerado la adopción de sistemas automatizados de toma de decisiones en áreas como la salud pública y la vigilancia, planteando nuevos desafíos éticos y subrayando la importancia de un enfoque ético en el desarrollo y despliegue de estas tecnologías.

Esta evolución histórica demuestra cómo la automatización en la toma de decisiones éticas ha pasado de ser un concepto teórico a una realidad práctica con implicaciones significativas para la sociedad, requiriendo una continua reflexión y adaptación de los marcos éticos y regulatorios.

Automatización moral entre ética algorítmica y humana

La automatización moral implica la interacción entre dos conceptos clave: la ética algorítmica y la ética humana. Esta se sitúa en la intersección de la ética algorítmica y la ética humana, dos campos que, aunque relacionados, presentan diferencias fundamentales en su enfoque y aplicación (Wallach, 2010). (Consulte Tabla 2).

Tabla 2: Automatización Moral: Cuadro Comparativo - Ética Algorítmica vs. Ética Humana

Característica	Ética Algorítmica	Ética Humana
Definición	Delegación de decisiones éticas a sistemas de inteligencia artificial (IA).	Capacidad de los individuos para tomar decisiones morales considerando situaciones complejas y múltiples perspectivas.
Fundamento	Programación de sistemas automatizados con reglas y criterios predefinidos.	Empatía, intuición, experiencia personal y capacidad de adaptación a nuevas circunstancias.
Fortalezas	- Precisión y consistencia. - Minimización de sesgos (en teoría).	- Flexibilidad para considerar una amplia gama de factores. - Adaptación al contexto.
Debilidades	- Falta de capacidad para interpretar el contexto y las complejidades emocionales. - Perpetuación de sesgos preexistentes (en la práctica).	- Susceptibilidad a sesgos cognitivos y emocionales. - Inconsistencias y errores.
Ejemplos	- Algoritmos de predicción de reincidencia en la justicia penal.	- Toma de decisiones médicas o financieras complejas.
Ventajas en la Automatización Moral	- Análisis de datos objetivos y consistentes. - Reducción del impacto de sesgos humanos.	- Consideración de valores sociales y morales aceptables.
Desafíos en la Automatización Moral	- Dificultad para incorporar principios éticos complejos en algoritmos. - Falta de transparencia y rendición de cuentas en los sistemas de IA.	- Necesidad de supervisión y ajuste humano constante.
Autores Citados	- Wallach, H. (2010). Moral Machines: Ethics and the Perfect Code. Oxford University Press.	- Angwin, J., Larson, J., Mattu, S., & Kirchner, K. (2016). Machine bias that harms. ACM SIGKDD Explorations, 16(1), 22-38.

Fuente: Elaboración propia

De la tabla se extrae como análisis que la ética algorítmica se basa en la programación de sistemas automatizados para que tomen decisiones siguiendo reglas y criterios predefinidos. Este enfoque se orienta a la creación de algoritmos que maximicen la eficiencia y minimicen el sesgo, integrando ciertos estándares éticos en su funcionamiento. Sin embargo, los sistemas algorítmicos, a pesar de su precisión y consistencia, carecen de la capacidad para interpretar el contexto y las complejidades emocionales inherentes a las decisiones morales humanas. Un ejemplo relevante es el uso de algoritmos en la justicia penal para predecir la probabilidad de reincidencia. Estos algoritmos, aunque basados en datos, han sido criticados por perpetuar sesgos preexistentes y por no tener en cuenta factores humanos importantes (Angwin, Larson, Mattu, & Kirchner, 2016).

Por otro lado, la ética humana se fundamenta en la capacidad de los individuos para considerar situaciones complejas y evaluar múltiples perspectivas a través de la empatía, la intuición y la experiencia personal. Los seres humanos pueden adaptarse a nuevas circunstancias y ajustar sus decisiones según el contexto. No obstante, la toma de decisiones humanas puede verse afectada por sesgos cognitivos y emocionales, lo que puede llevar a inconsistencias y

errores. La flexibilidad de la ética humana permite considerar una amplia gama de factores, pero esta misma flexibilidad puede resultar en decisiones menos uniformes (Floridi, 2019).

La interrelación entre la ética algorítmica y la ética humana es un área de creciente interés y debate. Por un lado, los algoritmos pueden complementar la toma de decisiones humanas al proporcionar análisis de datos objetivos y consistentes, ayudando a reducir el impacto de los sesgos humanos. Por otro lado, es fundamental que los diseñadores de algoritmos incorporen principios éticos humanos en el desarrollo de estos sistemas para garantizar que las decisiones automatizadas reflejen valores sociales y morales aceptables.

Comparar la ética algorítmica y la ética humana revela tanto fortalezas como debilidades en cada enfoque. La ética algorítmica ofrece consistencia y objetividad, pero puede carecer de la capacidad de adaptarse a contextos complejos y matizados. En contraste, la ética humana es flexible y capaz de considerar una amplia gama de factores, pero puede ser inconsistente y susceptible a sesgos. Un equilibrio entre ambos enfoques puede ser la clave para una automatización moral efectiva y ética.

Creencias en torno a la automatización moral

Las creencias en torno a la automatización moral reflejan un debate complejo sobre la capacidad de las tecnologías para tomar decisiones éticas de manera autónoma. Por un lado, existe una corriente de pensamiento que sostiene que los sistemas automatizados pueden procesar grandes cantidades de información y aplicar reglas lógicas de manera más eficiente y objetiva que los seres humanos. Esta perspectiva se fundamenta en la idea de que los algoritmos pueden ser programados para seguir principios éticos predefinidos sin verse afectados por sesgos emocionales o cognitivos.

Sin embargo, esta visión optimista de la automatización moral enfrenta críticas significativas. Expertos argumentan que la toma de decisiones éticas implica una complejidad que va más allá de la simple aplicación de reglas lógicas. Wallach y Allen (2008) señalan que las decisiones morales a menudo requieren una comprensión contextual y una capacidad de empatía que los sistemas automatizados actuales no poseen. Estos autores sostienen que la moralidad humana implica una interacción compleja entre razón, emoción y experiencia que es difícil de replicar en máquinas.

Un ejemplo concreto de las limitaciones de la automatización moral se observa en el ámbito de la justicia predictiva. Aunque estos sistemas se presentan como soluciones objetivas para reducir el sesgo en las decisiones judiciales, investigaciones han demostrado que pueden perpetuar y amplificar desigualdades existentes. De acuerdo, hay diversos estudios que han demostrado cómo los sistemas de IA utilizados en el ámbito judicial pueden perpetuar y amplificar las desigualdades existentes en la sociedad.

Una evidencia relevante es el caso del algoritmo COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), utilizado en Estados Unidos para evaluar el riesgo de reincidencia de los delincuentes. Un estudio realizado por la Universidad de Wharton encontró que este algoritmo tenía un sesgo racial significativo, siendo más propenso a clasificar erróneamente a los acusados afroamericanos como de alto riesgo en comparación con los acusados blancos (Angwin et al., 2016).

Otro ejemplo es la investigación realizada por el MIT Media Lab, que analizó los algoritmos utilizados en los sistemas de predicción de crímenes utilizados por la policía. Los resultados mostraron que estos sistemas tendían a predecir más delitos en áreas con mayor proporción de minorías raciales, perpetuando así los patrones de vigilancia y encarcelamiento desproporcionado de estas comunidades (Hao, 2019).

Estos hallazgos ponen en evidencia cómo los sesgos presentes en los datos utilizados para entrenar los algoritmos, así como en los propios procesos de diseño y desarrollo, pueden dar lugar a resultados discriminatorios y amplificar las desigualdades existentes, incluso cuando estos sistemas se presentan como soluciones objetivas y neutrales (Barocas & Selbst, 2016).

Es necesario que los desarrolladores de estos sistemas de IA y los responsables de su implementación en el ámbito judicial sean conscientes de estos riesgos y adopten medidas para mitigar los sesgos, mediante, por ejemplo, la diversificación de los equipos de desarrollo, el uso de técnicas de auditoría y evaluación de los algoritmos, y la transparencia en los procesos de toma de decisiones (Selbst et al., 2019).

Posición intermedia

Por otra parte, algunos investigadores adoptan una posición intermedia, sugiriendo que la automatización moral puede ser una herramienta valiosa cuando se combina con la supervisión y el juicio humano. Rueda, (2023) argumenta que la inteligencia artificial puede asistir en la toma de decisiones éticas, especialmente en situaciones que involucran grandes cantidades de datos o contextos complejos. Sin embargo, enfatiza la importancia de mantener la responsabilidad humana en el proceso de toma de decisiones.

Es fundamental reconocer que la programación de sistemas automatizados para la toma de decisiones éticas implica inevitablemente la incorporación de valores y juicios humanos. Como señala Mittelstadt et al. (2016), los algoritmos no son neutrales, sino que reflejan las prioridades y sesgos de sus creadores. Por lo tanto, la idea de una automatización moral completamente objetiva y libre de influencia humana es, en la práctica, inalcanzable.

Realidad detrás de las tecnologías automatizadas

Sin embargo, la realidad es que la automatización completa de decisiones éticas es una ilusión. A pesar de la apariencia de objetividad, las tecnologías automatizadas están diseñadas y programadas por seres humanos, quienes

introducen sus propios valores, sesgos y limitaciones en los algoritmos. Los sistemas de IA y los algoritmos operan basándose en los datos que reciben y las reglas que se les imponen, lo que significa que están intrínsecamente vinculados a los intereses y decisiones de quienes los desarrollan.

Por ejemplo, los datos utilizados para entrenar los algoritmos pueden contener sesgos históricos y sistemáticos que se reflejan en las decisiones de los sistemas automatizados. Esto puede dar lugar a resultados injustos o sesgados, a pesar de la intención de crear un sistema imparcial (O'Neil, C. (2016). Además, los algoritmos carecen de la capacidad para comprender el contexto complejo y las sutilezas morales que los humanos consideran en la toma de decisiones éticas (Binns, et all, 2018).

El interés humano detrás de la automatización

Detrás de la automatización de decisiones éticas, el interés humano sigue siendo fundamental. Las decisiones éticas no solo se basan en datos y reglas, sino que también involucran una comprensión profunda de los valores humanos, el contexto cultural y las implicaciones sociales. Los desarrolladores de tecnologías automatizadas toman decisiones sobre cómo estructurar los algoritmos, qué datos utilizar y qué principios éticos implementar (Cornelio, Rodríguez et al. 2024). Por lo tanto, los sistemas automatizados están imbuidos de los intereses y perspectivas de quienes los diseñan (Dastin, 2018).

La responsabilidad última de las decisiones éticas siempre recae en los seres humanos (Arencibia, Cornelio et al. 2024), (Arencibia, Cornelio et al. 2024). Aunque los sistemas automatizados pueden proporcionar recomendaciones y soporte en la toma de decisiones, la interpretación y aplicación final de estas decisiones deben ser realizadas por humanos, quienes tienen la capacidad de considerar el impacto ético y social de sus acciones (Wallach, & Allen, 2009).

Aunque las tecnologías automatizadas pueden ofrecer soluciones eficientes para la toma de decisiones, no se les pueden delegar por completo la responsabilidad ética. La automatización de decisiones éticas está intrínsecamente vinculada a los intereses humanos, tanto en su diseño como en su aplicación. Reconocer y abordar estas limitaciones es esencial para desarrollar sistemas de automatización que respeten los principios éticos y contribuyan al bienestar social de manera justa y equitativa.

Desafíos de la automatización moral

La automatización moral enfrenta diversos desafíos y limitaciones que requieren un análisis cuidadoso para su implementación efectiva y ética. Estos obstáculos abarcan desde la complejidad de codificar valores éticos hasta las implicaciones de la responsabilidad y la rendición de cuentas en sistemas automatizados.

Uno de los principales desafíos en la automatización moral es la codificación de valores éticos en los algoritmos. A diferencia de los humanos, que pueden adaptarse y aplicar juicios morales basados en contextos y experiencias cambiantes, los algoritmos requieren reglas y parámetros predefinidos (Binns et al., 2018). La dificultad radica en traducir principios éticos complejos y a menudo subjetivos en códigos lógicos y objetivos. Este proceso puede resultar en la simplificación excesiva de valores morales, lo que puede llevar a decisiones que no reflejan adecuadamente las complejidades del juicio humano.

El manejo de situaciones ambiguas y dilemas morales constituye otro desafío importante. Los sistemas automatizados, a pesar de su capacidad para procesar grandes cantidades de datos, carecen de la intuición y el juicio contextual que los seres humanos aplican en situaciones moralmente complejas. Un estudio realizado por Awad et al. (2018) sobre dilemas morales en vehículos autónomos reveló que las preferencias éticas varían significativamente entre culturas, lo que complica aún más la implementación de un sistema ético universal. La capacidad limitada de los algoritmos para entender y evaluar contextos complejos puede resultar en decisiones que no coinciden con las expectativas éticas humanas.

La transparencia y comprensibilidad de los algoritmos es fundamental para generar confianza en los sistemas de automatización moral. Sin embargo, la complejidad de los algoritmos de aprendizaje automático avanzados puede hacer que sus procesos de toma de decisiones sean opacos incluso para sus creadores. Esta "caja negra" algorítmica plantea preocupaciones sobre la capacidad de auditar y comprender cómo se toman las decisiones éticas automatizadas (Burrell, 2016). Esta falta de transparencia puede dificultar la identificación de errores y sesgos en el proceso de toma de decisiones, y socavar la confianza del público en la automatización moral. La necesidad de algoritmos interpretables y transparentes es fundamental para garantizar la rendición de cuentas y la confianza en las decisiones automatizadas.

Los sesgos inherentes y la discriminación algorítmica representan una preocupación significativa en la automatización moral. Los algoritmos entrenados en datos históricos pueden perpetuar o incluso exacerbar los sesgos existentes en esos datos (Angwin et al., 2016). Por ejemplo, los sistemas de justicia predictiva han sido criticados por sesgos raciales y socioeconómicos, lo que lleva a decisiones injustas y discriminatorias. Es esencial desarrollar técnicas para identificar y mitigar los sesgos en los algoritmos para garantizar la equidad y la justicia en la toma de decisiones automatizada.

La responsabilidad y rendición de cuentas en sistemas automatizados plantea desafíos legales y éticos complejos. Determinar quién es responsable cuando un sistema automatizado toma una decisión ética incorrecta o dañina no es sencillo. Mittelstadt et al. (2016) argumentan que la falta de un marco claro para la atribución de responsabilidad en

sistemas de IA puede obstaculizar su adopción y aceptación social. Los diseñadores, programadores y operadores de estos sistemas deben asumir la responsabilidad de garantizar que los algoritmos funcionen de manera ética y justa. Además, deben establecerse mecanismos claros de rendición de cuentas para abordar las fallas y errores en los sistemas automatizados.

Estos desafíos y limitaciones subrayan la necesidad de un enfoque multidisciplinario en el desarrollo de sistemas de automatización moral. La colaboración entre expertos en ética, tecnología, derecho y ciencias sociales es esencial para abordar estas complejidades y desarrollar soluciones que equilibren la eficiencia de la automatización con los valores éticos fundamentales de la sociedad.

Consideraciones éticas

La automatización moral plantea profundas consideraciones éticas que los límites de la automatización moral (Consulte Tabla 3). Este campo de estudio examina cómo la implementación de sistemas automatizados para la toma de decisiones éticas afecta la autonomía humana, la responsabilidad moral y los fundamentos mismos de la ética.

Tabla 3: Límites de la automatización moral.

Limitación	Descripción
Falta de inteligencia emocional y contextualización cultural	- Incapacidad de captar sutilezas emocionales y culturales. - Decisiones técnicamente correctas pero inapropiadas o insensibles en ciertos contextos.
Simplificación excesiva de dilemas morales complejos	- Operan con base en reglas predefinidas y datos históricos. - Decisiones que no reflejan adecuadamente la complejidad de los problemas morales.
Perpetuación y amplificación de sesgos	- Entrenamiento con datos sesgados o con prejuicios históricos. - Decisiones discriminatorias o injustas.
Falta de transparencia y comprensibilidad	- Dificultad para auditar y comprender cómo se toman las decisiones. - Erosión de la confianza pública y dificultad para identificar errores o sesgos.
Dificultad para determinar la responsabilidad	- Complejidad para atribuir la responsabilidad en caso de decisiones incorrectas o dañinas. - Erosión de la confianza en los sistemas y organizaciones que los emplean.
Incapacidad de comprender marcos éticos complejos	- Falta de comprensión profunda de las circunstancias y valores humanos. - Dificultad para tomar decisiones éticas adecuadas.
Necesidad de intervención humana en situaciones complejas	- Juicio ético complejo, empatía y comprensión contextual. - Campos como la atención médica y la justicia penal requieren supervisión humana.
Riesgo de erosión de la autonomía humana	- Dependencia excesiva de la automatización. - Reducción de la capacidad de tomar decisiones basadas en juicios y valores propios.

Fuente: Elaboración propia

De la Tabla se puede extraer como conclusión que los límites de la automatización moral se vuelven particularmente evidentes cuando la tecnología se utiliza de manera inadecuada, y es incapaz de incorporar las intuiciones humanas, llevando a distorsiones éticas significativas. Estos límites reflejan las restricciones inherentes de los sistemas automatizados y destacan la necesidad de supervisión humana para evitar consecuencias no deseadas. Por ejemplo, delegar decisiones éticas a algoritmos plantea serias preocupaciones sobre la capacidad de estas tecnologías para manejar la complejidad y la sutileza de las decisiones morales.

La ética algorítmica, aunque puede ser eficiente y objetiva en ciertos aspectos, a menudo carece de la profundidad y la comprensión contextual que caracterizan al juicio humano. Binns et al. (2018) argumentan que la programación de algoritmos con valores éticos predefinidos puede llevar a una simplificación excesiva de cuestiones morales complejas, lo que puede resultar en decisiones éticamente problemáticas. La falta de comprensión contextual y de empatía en los algoritmos puede provocar respuestas inapropiadas en situaciones que requieren una evaluación moral más matizada.

La automatización de decisiones éticas también impacta la autonomía humana y la agencia moral. Confiar en algoritmos para tomar decisiones puede reducir la responsabilidad personal y la capacidad de los individuos para tomar decisiones basadas en sus propios juicios y valores. Floridi (2019) señala que la creciente dependencia de la automatización puede erosionar la capacidad de las personas para actuar como agentes morales autónomos, debilitando su sentido de responsabilidad y control sobre las decisiones éticas. Coeckelbergh (2020) advierte que esta delegación puede llevar a una "atrofia moral", en la que las personas pierden la habilidad de razonar éticamente por sí mismas, lo que podría tener consecuencias a largo plazo en la formación del juicio moral individual y colectivo.

La supervisión humana es esencial para garantizar que las decisiones tomadas por algoritmos sean éticamente aceptables y estén alineadas con los valores sociales. Pasquale (2015) sostiene que la supervisión humana puede proporcionar una capa adicional de evaluación crítica, asegurando que las decisiones automatizadas sean revisadas y ajustadas según sea necesario para reflejar consideraciones éticas más amplias. Esta supervisión es importante para mantener la transparencia y la rendición de cuentas en la automatización moral. Mittelstadt et al. (2016) subrayan la importancia de mantener un "humano en el circuito" para supervisar y, cuando sea necesario, anular las decisiones de los sistemas automatizados. La supervisión humana es especialmente relevante en áreas sensibles como la justicia penal o la atención médica, donde las consecuencias de las decisiones éticas pueden tener un impacto significativo en la vida de las personas.

Los límites filosóficos y conceptuales de la automatización moral se manifiestan en el problema del marco en ética, que se refiere a la dificultad de especificar completamente el contexto relevante para una decisión ética. Dennett (1984) argumenta que este problema es particularmente desafiante para los sistemas artificiales, que carecen de la comprensión intuitiva del mundo que poseen los humanos. Esta falta de intuición y comprensión contextual puede limitar la capacidad de los algoritmos para tomar decisiones éticas que reflejen fielmente los valores humanos.

La cuestión de la conciencia y la intencionalidad también plantea límites conceptuales significativos. Searle (1980), a través de su experimento mental de la "habitación china", cuestiona la capacidad de los sistemas artificiales para comprender verdaderamente el significado de sus acciones, un componente que muchos consideran esencial para la

agencia moral. La falta de conciencia y de intencionalidad en los sistemas automatizados sugiere que, aunque puedan simular decisiones éticas, no pueden comprender plenamente las implicaciones morales de sus acciones.

El impacto en la responsabilidad y la agencia moral es otro aspecto crítico. La automatización de decisiones éticas puede difuminar las líneas de responsabilidad, planteando preguntas sobre quién es responsable cuando un sistema automatizado toma una decisión éticamente cuestionable. Este dilema se ha manifestado en casos como el accidente fatal de un vehículo autónomo de Uber en 2018, que planteó preguntas complejas sobre la responsabilidad legal y moral. La dificultad para atribuir responsabilidad en tales casos puede socavar la confianza pública en la tecnología y en los sistemas que la utilizan.

La automatización moral, aunque promete eficiencia y consistencia en la toma de decisiones éticas, enfrenta desafíos significativos que requieren una consideración cuidadosa. La integración de estos sistemas debe equilibrar los beneficios de la automatización con la preservación de la autonomía humana y la responsabilidad moral. El desarrollo futuro de la automatización moral necesitará un enfoque interdisciplinario que aborde tanto los aspectos técnicos como los filosóficos y sociales de la ética computacional. Es esencial que los diseñadores y reguladores de estas tecnologías trabajen juntos para desarrollar marcos éticos que guíen el uso de la automatización en decisiones morales, garantizando que estas tecnologías sean utilizadas de manera responsable y justa.

Es necesario entender que detrás de cada sistema automatizado existe una conciencia humana que lo diseña, programa y supervisa. Esta realidad subraya la importancia de mantener altos estándares éticos en todo el proceso de desarrollo e implementación de estos sistemas. La integridad y responsabilidad de los individuos involucrados en la creación de sistemas de automatización moral son fundamentales para garantizar que estos operen de manera ética y beneficiosa para la sociedad.

Una de las principales limitaciones de la automatización moral es su carencia de inteligencia emocional y contextualización cultural. Los sistemas automatizados, por avanzados que sean, carecen de la capacidad de comprender las sutilezas emocionales y culturales que a menudo son cruciales en la toma de decisiones éticas. Esta falta de comprensión contextual puede llevar a decisiones que, aunque lógicamente correctas según los parámetros programados, pueden ser inapropiadas o incluso perjudiciales en ciertos contextos culturales o emocionales.

La dependencia de estos sistemas en recursos externos, como la electricidad y el mantenimiento técnico, subraya su falta de autonomía real. Sin la intervención y supervisión humana continua, los sistemas de automatización moral no pueden funcionar ni adaptarse a nuevas situaciones o cambios en el entorno ético. Esta dependencia plantea preguntas sobre la fiabilidad y la aplicabilidad de estos sistemas en situaciones críticas o de emergencia donde la toma de decisiones éticas rápida y autónoma puede ser necesaria.

La transparencia y comprensibilidad de los algoritmos utilizados en la automatización moral es otro aspecto que merece una consideración cuidadosa. La complejidad de muchos sistemas de inteligencia artificial hace que sus procesos de toma de decisiones sean opacos, lo que dificulta la auditoría y la corrección de errores o sesgos. Esta falta de transparencia puede socavar la confianza pública en estos sistemas y plantear serios desafíos éticos y legales.

Los sesgos inherentes y la posibilidad de discriminación algorítmica son preocupaciones significativas. Los sistemas automatizados, al ser entrenados con datos históricos, pueden perpetuar y amplificar prejuicios existentes en la sociedad. Esto plantea el riesgo de que las decisiones automatizadas refuercen desigualdades y discriminaciones, en lugar de mitigarlas.

La cuestión de la responsabilidad y la rendición de cuentas en la toma de decisiones automatizadas es un área que requiere una consideración profunda. Determinar quién es responsable cuando un sistema automatizado toma una decisión ética incorrecta o dañina es un desafío complejo que tiene implicaciones legales y morales significativas.

Desde una perspectiva filosófica, la automatización moral enfrenta limitaciones conceptuales importantes. La ausencia de conciencia y intencionalidad en estos sistemas plantea preguntas fundamentales sobre su capacidad para comprender verdaderamente las implicaciones éticas de sus decisiones. Además, la complejidad de muchos dilemas éticos va más allá de lo que se puede codificar en reglas lógicas, lo que pone en duda la capacidad de estos sistemas para abordar adecuadamente situaciones morales complejas.

Las implicaciones sociales y culturales de la automatización moral son amplias y variadas. La diversidad ética y cultural presente en las sociedades humanas presenta un desafío significativo para la implementación de sistemas éticos universales. Lo que se considera ético en una cultura puede no serlo en otra, lo que complica la creación de sistemas de automatización moral que sean aplicables y aceptables globalmente.

Análisis comparativo de casos

Comparar las decisiones algorítmicas y humanas en contextos específicos revela importantes diferencias en la capacidad para manejar la complejidad ética. Las decisiones humanas, con su capacidad para empatía e intuición, contrastan con las decisiones algorítmicas, que a menudo se basan en parámetros predefinidos y datos históricos.

En el ámbito financiero, un ejemplo notable de éxito en la automatización moral es el uso de algoritmos en la detección de fraude. Las instituciones financieras han implementado sistemas automatizados que analizan patrones en transacciones para identificar actividades sospechosas. Estos sistemas han mostrado una gran efectividad en la reducción de fraudes, detectando patrones que pueden ser difíciles de identificar manualmente (FICO, 2018). Este éxito demuestra la capacidad de los algoritmos para manejar grandes volúmenes de datos y realizar análisis detallados que pueden superar la capacidad humana en términos de velocidad y precisión.

Sin embargo, la automatización moral también ha enfrentado fracasos significativos. Un caso relevante es el de la plataforma de contratación de Amazon que, en 2018, desechó un sistema de contratación automatizado debido a sesgos de género. El algoritmo, diseñado para analizar currículos y seleccionar candidatos, mostró una preferencia sistemática por candidatos masculinos, en parte debido a que fue entrenado con datos históricos que reflejaban una predominancia de hombres en puestos tecnológicos (Dastin, 2018). Este caso ilustra cómo los algoritmos pueden perpetuar y amplificar los sesgos existentes en los datos, lo que lleva a resultados no éticos que no reflejan una verdadera igualdad de oportunidades.

Las lecciones aprendidas de estos casos destacan la necesidad de una supervisión humana constante y un enfoque proactivo para abordar los sesgos en los sistemas automatizados. La transparencia en el diseño y la implementación de algoritmos es crucial para garantizar la equidad. Pasquale (2015) subraya la importancia de la supervisión y la rendición de cuentas para que las decisiones automatizadas sean justas y equitativas. Mittelstadt et al. (2016) enfatizan la necesidad de mantener un "humano en el circuito" para monitorear y ajustar las decisiones automatizadas, asegurando que estas decisiones se alineen con valores éticos más amplios y contextos específicos.

Perspectivas futuras y recomendaciones

Las perspectivas futuras en el campo de la ética algorítmica y la automatización moral están marcadas por la evolución tecnológica, la necesidad de una regulación ética adecuada y la integración de principios éticos humanos en el diseño de algoritmos.

Innovaciones tecnológicas y su impacto en la ética algorítmica

Las innovaciones tecnológicas están transformando rápidamente el paisaje de la ética algorítmica. Avances en inteligencia artificial y aprendizaje automático están mejorando la capacidad de los algoritmos para analizar grandes volúmenes de datos y realizar predicciones precisas. Sin embargo, estos avances también plantean nuevos desafíos éticos. La implementación de tecnologías emergentes, como el aprendizaje profundo y los sistemas de IA explicativos, tiene el potencial de mejorar la transparencia y la comprensión de las decisiones algorítmicas (Samek et al., 2019). Estas tecnologías pueden ofrecer una mayor capacidad para desentrañar los procesos internos de los algoritmos y hacer que sus decisiones sean más comprensibles para los humanos.

No obstante, la integración de estas tecnologías también requiere una consideración cuidadosa de sus implicaciones éticas. Por ejemplo, los sistemas de IA explicativos deben ser diseñados para evitar la sobreexposición de datos sensibles y garantizar la protección de la privacidad (Doshi-Velez & Kim, 2017). La continua evolución de la tecnología subraya la necesidad de adaptar las estrategias éticas para abordar los desafíos emergentes y garantizar que los avances tecnológicos se utilicen de manera justa y responsable.

Propuestas para una regulación ética de algoritmos

La regulación ética de los algoritmos es esencial para garantizar que las tecnologías automatizadas sean utilizadas de manera equitativa y justa. Propuestas para una regulación efectiva incluyen la creación de marcos normativos que obliguen a las organizaciones a realizar evaluaciones de impacto ético antes de implementar sistemas automatizados (Floridi, 2019). Estas evaluaciones deben abordar no solo los riesgos potenciales de los algoritmos, sino también las medidas para mitigar posibles sesgos y garantizar la transparencia en los procesos de toma de decisiones.

Además, la implementación de principios de "diseño ético por defecto" es fundamental. Esto implica incorporar consideraciones éticas desde las etapas iniciales del desarrollo de algoritmos, asegurando que se integren mecanismos para detectar y corregir sesgos, así como para proteger los derechos de los individuos afectados (O'Neil, 2016). La regulación también debe contemplar la creación de órganos independientes de supervisión para auditar y revisar los sistemas automatizados, garantizando su conformidad con los principios éticos y legales establecidos.

Integración de la ética humana en el diseño de algoritmos

Integrar la ética humana en el diseño de algoritmos es una estrategia clave para abordar los desafíos éticos de la automatización moral. Esto implica no solo incluir principios éticos en el diseño de algoritmos, sino también garantizar que estos sistemas reflejen una comprensión profunda de los valores y normas sociales (Dastin, 2018). La participación de expertos en ética y representantes de la sociedad civil en el proceso de desarrollo de algoritmos puede ayudar a asegurar que estos sistemas sean diseñados para abordar las necesidades y preocupaciones de diversas comunidades.

Asimismo, es importante fomentar la educación y la capacitación en ética para los desarrolladores de algoritmos. La formación en principios éticos y en la identificación y mitigación de sesgos puede preparar mejor a los profesionales para enfrentar los desafíos éticos en el desarrollo de tecnologías automatizadas (Mittelstadt et al., 2016). La integración de la ética en el diseño de algoritmos no solo mejora la equidad y la justicia en los sistemas automatizados, sino que también contribuye a la confianza pública en estas tecnologías.

Discusión

La discusión sobre los límites de la automatización moral revela la complejidad de aplicar principios éticos en sistemas de inteligencia artificial (IA). Aunque la teoría de la ética algorítmica sugiere que los sistemas automatizados pueden incorporar valores morales, en la práctica, estos sistemas a menudo reflejan y amplifican los sesgos humanos existentes. Los estudios revisados, como el de Awad et al. (2018) y los casos de sesgos en herramientas de contratación y algoritmos predictivos, demuestran que la automatización puede perpetuar desigualdades en lugar de

resolverlas. Este desafío subraya la dificultad de codificar valores éticos en algoritmos de manera que sean justos y equitativos.

A pesar de los esfuerzos por regular y guiar el desarrollo de IA, como se evidencia en la Ley de Inteligencia Artificial de la Comisión Europea, las regulaciones actuales parecen insuficientes para abordar los problemas éticos fundamentales que surgen con la automatización. Los marcos regulatorios y las políticas propuestas a menudo no capturan la complejidad y la dinámica de los sistemas algorítmicos en uso, lo que limita su eficacia. Esto plantea la necesidad de una revisión más profunda de las políticas y la implementación de normas más estrictas que consideren las implicaciones éticas en contextos específicos y variados.

Conclusiones

El análisis de los límites de la automatización moral subraya que, aunque los algoritmos pueden mejorar la eficiencia y consistencia en la toma de decisiones, no son adecuados para todas las situaciones, especialmente aquellas que requieren matices emocionales, comprensión cultural y empatía. En áreas como la atención médica y la justicia penal, la intervención humana sigue siendo esencial para abordar la complejidad y sensibilidad de las decisiones éticas.

La responsabilidad humana es crucial en el uso de sistemas automatizados. Es necesario mantener una supervisión constante para asegurar que las decisiones algorítmicas se alineen con los valores éticos, y los desarrolladores deben ser conscientes de los impactos sociales y éticos de sus algoritmos. La transparencia y la rendición de cuentas son fundamentales para mantener la confianza pública y garantizar un uso justo de estas tecnologías.

Para establecer límites efectivos a la automatización moral, es vital implementar mecanismos de auditoría detallada que incluyan documentación clara de datos, criterios y métodos. La colaboración entre desarrolladores, expertos en ética y la sociedad en general puede fomentar una mayor transparencia y responsabilidad, asegurando que la automatización respete los principios de justicia y empatía.

Conflictos de intereses

El autor no poseen conflictos de intereses.

Contribución de los autores

1. Conceptualización: Mario González Arencibia.
2. Curación de datos: Mario González Arencibia.
3. Análisis formal: Mario González Arencibia, Omar Mar Cornelio.
4. Investigación: Mario González Arencibia, Omar Mar Cornelio.
5. Metodología: Mario González Arencibia, Omar Mar Cornelio.
6. Software: Mario González Arencibia.
7. Validación: Mario González Arencibia, Omar Mar Cornelio.
8. Visualización: Mario González Arencibia, Omar Mar Cornelio.
9. Redacción – borrador original: Mario González Arencibia, Omar Mar Cornelio.
10. Redacción – revisión y edición: Mario González Arencibia.

Financiamiento

La investigación no requirió fuente de financiamiento externa.

Referencias

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The Moral Machine experiment. Nature, 563(7729), 59-64.

Arencibia, M. G., et al. (2024). "Aspectos éticos de la aplicación de la informática a la medicina." Serie Científica de la Universidad de las Ciencias Informáticas 17(8): 1-18.

Arencibia, M. G., et al. (2024). "Ética digital en la salud." Serie Científica de la Universidad de las Ciencias Informáticas 17(5): 22-39.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104, 671-732. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). *The Role of Moral Values in the Design and Deployment of AI Systems*. *Proceedings of the 2018 CHI Conference*

Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company.

Bryson, J. J. (2019). The Past Decade and the Future of AI Ethics. Nature Machine Intelligence, 1(1), 5-6. Recuperado de: https://www.nature.com/articles/s42256-018-0025-2

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 2053951715622512.

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian.

Coeckelbergh, M. (2020). *AI Ethics*. The MIT Press.

Cornelio, O. M., et al. (2024). "La Inteligencia Artificial: desafíos para la educación." Editorial Internacional Alema.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from Reuters.

Dennett, D. C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press.

Diakopoulos, N. (2016). Algorithmic accountability: Journalistic investigations of computational power structures. *Digital Journalism, 4*(2), 550-565. Retrieved from Digital Journalism.

Doshi-Velez, F., & Kim, P. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. arXiv.

European Commission. (2021). *Artificial Intelligence Act*. Available at: ec.europa.eu.

European Union. (2016). General Data Protection Regulation. Official Journal of the European Union, L119, 1-88.

FICO. (2018). *The Future of Fraud Detection: How Artificial Intelligence is Changing the Game*. Retrieved from FICO.

Floridi, L. (2019). The ethics of artificial intelligence. In *The Cambridge Handbook of Information and Computer Ethics* (pp. 116-134). Cambridge University Press. Retrieved from Cambridge University Press.

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. Minds and Machines, 14(3), 349-379. Recuperado de: https://link.springer.com/article/10.1023/B:MIND.0000035461.63578.9d

Haidt, J. (2012). The Righteous Mind: Why Good People Are Divided by Politics and Religion. Pantheon.

Hao, K. (2019). Predictive Policing Algorithms Are Racist. That Doesn't Mean They'll Stop Using Them. MIT Technology Review. Recuperado de: https://www.technologyreview.com/2019/07/26/612/predictive-policing-algorithms-racist-police-criminal-justice/

IEEE. (2020). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Artificial Intelligence and Autonomous Systems*. Available at: ieee.org.

Kohlberg, L. (1981). The Philosophy of Moral Development: Moral Stages and the Idea of Justice. Harper & Row.

Metzinger, T. (2019). Ethics Washing Made in Europe. Der Tagesspiegel. Recuperado de: https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

Moor, J. H. (1985). What is Computer Ethics? Metaphilosophy, 16(4), 266-275.

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. IEEE Intelligent Systems, 21(4), 18-21. Recuperado de: https://ieeexplore.ieee.org/document/1667948

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Rueda, J. (2023). ¿Automatizando la mejora moral humana? La inteligencia artificial para la ética: Nota crítica sobre LARA, F. y J. SAVULESCU (eds.) (2021), Más (que) humanos. Biotecnología, inteligencia artificial y ética de la mejora. Madrid: Tecnos. *Daimon Revista Internacional de Filosofia*, (89), 199–209. https://doi.org/10.6018/daimon.508771

Samek, W., Wiegand, T., & Müller, K. R. (2019). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITuU Journal*, 2(1), 68-77. Retrieved from Springer.

Searle, J. R. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences*, 3(3), 417-457. https://doi.org/10.1017/S0140525X00005756

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59-68. https://dl.acm.org/doi/10.1145/3287560.3287598

Turing, A. M. (1950). Computing Machinery and Intelligence. Mind, 59(236), 433-460.

Wallach, W. (2010). *Moral Machines: Teaching Robots Right From Wrong*. Oxford University Press.

Weizenbaum, J. (1976). Computer Power and Human Reason: From Judgment to Calculation. W. H. Freeman & Co.