Revista UNESUM-Ciencias
Volumen 8, Número 2, 2024
Universidad Estatal del Sur de Manabí
ISSN-e: 2602-8166

ARTÍCULO ORIGINAL

Encrucijada entre Prompts éticos y juicio humano en la era de la Inteligencia Artificial

Crossroads between Ethical Prompts and Human Judgment in the Age of Artificial Intelligence

https://doi.org/10.47230/unesum-ciencias.v8.n2.2024.4-19

Recibido: 15-01-2024 **Aceptado:** 11-02-2024 **Publicado:** 20-05-2024

Mario González Arencibia[1*]
https://orcid.org/0000-0001-9947-7762

1. Centro de Estudios de Gestión de Proyectos y Toma de Decisiones; Facultad de Informática Organizacional; Universidad de Ciencias Informáticas; La Habana, Cuba.

Volumen: 8
Número: 2
Año: 2024
Paginación: 4-19
URL: https://revistas.unesum.edu.ec/index.php/unesumciencias/article/view/795
***Correspondencia autor:** mgarencibia@uci.cu

RESUMEN

En la era actual de rápido avance tecnológico, la inteligencia artificial (IA) ha penetrado en diversos aspectos de la toma de decisiones, incluyendo aquellas con implicaciones éticas significativas. Este estudio examina la compleja relación entre los prompts éticos implementados en sistemas de IA y el juicio moral humano, investigando si estos prompts pueden reemplazar efectivamente la toma de decisiones éticas humanas. El objetivo principal es evaluar las capacidades y limitaciones de los prompts éticos en la captura de la complejidad del razonamiento moral humano. La investigación revela que los prompts éticos han demostrado eficacia en situaciones donde las reglas éticas son claras y bien definidas, como en ciertos aspectos de la medicina y las finanzas, logrando en ocasiones igualar o superar el desempeño humano. Sin embargo, también se identificaron limitaciones significativas, particularmente en escenarios éticos complejos que requieren una comprensión contextual profunda y flexibilidad cognitiva. Los sistemas de IA guiados por prompts éticos mostraron dificultades para manejar variaciones sutiles en dilemas morales y para adaptarse a diferentes contextos culturales. La investigación concluye que, si bien los prompts éticos representan un avance importante en la incorporación de consideraciones éticas en sistemas de IA, no pueden reemplazar completamente el juicio humano en la toma de decisiones éticas complejas. Se sugiere un enfoque híbrido que combine las fortalezas de los prompts éticos con la supervisión y el juicio humano continuo como el camino más prometedor para abordar los desafíos éticos en la era de la IA.

Palabras clave: Inteligencia Artificial, Ética, Prompts Éticos, Juicio Moral, Toma de Decisiones, Razonamiento Contextual.

ABSTRACT

In the current era of rapid technological advancement, artificial intelligence (AI) has penetrated various aspects of decision-making, including those with significant ethical implications. This study examines the complex relationship between ethical prompts implemented in AI systems and human moral judgment, investigating whether these prompts can effectively replace human ethical decision-making. The main objective is to evaluate the capabilities and limitations of ethical prompts in capturing the complexity of human moral reasoning. The research reveals that ethical prompts have demonstrated effectiveness in situations where ethical rules are clear and well defined, such as in certain aspects of medicine and finance, sometimes matching or surpassing human performance. However, significant limitations were also identified, particularly in complex ethical scenarios that require deep contextual understanding and cognitive flexibility. AI systems guided by ethical prompts showed difficulties in handling subtle variations in moral dilemmas and adapting to different cultural contexts. The research concludes that, while ethical prompts represent an important advance in incorporating ethical considerations into AI systems, they cannot completely replace human judgment in complex ethical decision-making. A hybrid approach that combines the strengths of ethical prompts with continuous human supervision and judgment is suggested as the most promising path to address ethical challenges in the AI era.

Keywords: Artificial Intelligence, Ethics, Ethical Prompts, Moral Judgment, Decision Making, Contextual Reasoning.

Introducción

En la era digital actual, donde la inteligencia artificial (IA) se integra cada vez más en la toma de decisiones éticas, surge una pregunta fundamental: ¿Pueden los prompts éticos reemplazar el juicio humano en la codificación moral de los sistemas de IA? Esta cuestión se ha vuelto particularmente relevante a medida que las tecnologías de IA se despliegan en áreas sensibles como la atención médica, el sistema judicial y la seguridad pública.

El concepto de prompts éticos en IA se refiere a las instrucciones o directrices programadas para guiar el comportamiento ético de los sistemas inteligentes. Estos prompts buscan incorporar principios morales y valores humanos en el proceso de toma de decisiones de la IA. Sin embargo, la complejidad de las situaciones éticas del mundo real plantea desafíos significativos para la codificación de la moralidad en sistemas automatizados.

Estudios recientes han explorado la eficacia de los prompts éticos en la IA. Por ejemplo, una investigación realizada por Hagendorff (2020) examinó las limitaciones de los enfoques basados en reglas para la ética en IA, destacando la necesidad de considerar el contexto y la ambigüedad en situaciones éticas complejas. Asimismo, Bender et al. (2021) analizaron los riesgos éticos asociados con los modelos de lenguaje a gran escala, subrayando la importancia de la supervisión humana en la interpretación y aplicación de principios éticos.

La UNESCO han reconocido la importancia de abordar las implicaciones éticas de la IA. En 2021, la UNESCO adoptó una recomendación sobre la ética de la IA, enfatizando la necesidad de un enfoque centrado en el ser humano para el desarrollo y uso de estas tecnologías. Por su parte, empresas tecnológicas como Google y Microsoft han establecido comités de ética en IA, reconociendo la complejidad de implementar principios éticos en sistemas automatizados.

A pesar de estos avances, existe un vacío de conocimiento significativo en cuanto a la capacidad de los prompts éticos para capturar la totalidad del razonamiento moral humano. La naturaleza dinámica y contextual de muchas decisiones éticas plantea desafíos para la codificación de principios morales en sistemas de IA. Este estudio busca abordar esta brecha, explorando los límites y posibilidades de los prompts éticos en la toma de decisiones morales automatizadas.

Las preguntas centrales que guían esta investigación incluyen: ¿En qué medida pueden los prompts éticos capturar la complejidad del razonamiento moral humano? ¿Cuáles son los límites de la codificación ética en sistemas de IA? ¿Qué papel debe desempeñar la supervisión humana en la toma de decisiones éticas de la IA?

El objetivo principal de este estudio es examinar si los prompts éticos pueden sustituir el juicio humano en la toma de decisiones, evaluando tanto sus potencialidades como sus limitaciones. La importancia de esta investigación radica en sus implicaciones para el futuro desarrollo y despliegue de sistemas de IA éticos. Comprender las capacidades y limitaciones de los prompts éticos es esencial para diseñar sistemas de IA que puedan operar de manera segura y responsable en entornos complejos. Por otro lado, este estudio contribuye al debate sobre el papel de la IA en la sociedad y la medida en que podemos se debe confiar en sistemas automatizados para tomar decisiones éticas.

Código moral, prompts éticos, juicio humano, toma de decisiones éticas

En el contexto de la inteligencia artificial y la ética, es fundamental definir claramente los conceptos clave que subyacen a esta discusión. El código moral se refiere al conjunto de principios y valores que guían el comportamiento ético, ya sea de individuos o sistemas. En el ámbito de la IA, el código moral se traduce en reglas y directrices programables que buscan replicar o emular los estándares éticos humanos.

Mientras que el juicio humano, por su parte, representa la capacidad única de los seres humanos para evaluar situaciones complejas, considerando múltiples factores, contextos y consecuencias antes de llegar a una conclusión ética. Este proceso involucra no solo el conocimiento de principios morales, sino también la intuición, la empatía y la experiencia acumulada.

La toma de decisiones éticas se refiere al proceso de seleccionar un curso de acción entre varias alternativas, basándose en consideraciones morales y valores éticos. Este proceso implica la evaluación de las consecuencias potenciales, la ponderación de diferentes principios morales y la consideración del contexto específico en el que se toma la decisión.

La intersección entre el código moral, los prompts éticos y el juicio humano se ha convertido en un campo de estudio fascinante y complejo, en la era de la inteligencia artificial. Los prompts éticos, instrucciones diseñadas para guiar el comportamiento ético de los sistemas de IA, representan un intento de codificar principios morales en lenguaje máquina. Sin embargo, la pregunta persiste: ¿pueden estos prompts reemplazar genuinamente el juicio humano en la toma de decisiones éticas?

El filósofo Peter Singer (2021) argumenta que la ética no es simplemente un conjunto de reglas fijas, sino un proceso dinámico de razonamiento que requiere contextualización y adaptabilidad. En su libro "Ethics in the Real World: 82 Brief Essays on Things That Matter" (Princeton University Press, EE.UU.), Singer explora cómo los dilemas éticos a menudo involucran matices que los sistemas basados en reglas pueden pasar por alto. Esta perspectiva sugiere que los prompts éticos, por muy sofisticados que sean, podrían enfrentar limitaciones inherentes al abordar situaciones moralmente ambiguas.

Por otro lado, Stuart Russell (2019), en su obra "Human Compatible: Artificial Intelligence and the Problem of Control" (Viking, EE.UU.), propone que los sistemas de IA podrían, en teoría, ser diseñados para alinear sus objetivos con los valores humanos. Russell sugiere que, mediante un proceso de aprendizaje y refinamiento continuo, los prompts éticos podrían evolucionar para capturar aspectos más sutiles del razonamiento moral humano.

La interrelación entre estos conceptos se manifiesta en escenarios prácticos. Por ejemplo, en el campo de la medicina, los sistemas de IA ya están siendo utilizados para asistir en diagnósticos y planes de tratamiento. Un estudio realizado por Char et al. (2018) en el New England Journal of Medicine examinó los desafíos éticos de implementar algoritmos de aprendizaje automático en la atención médica. Los investigadores encontraron que, si bien estos sistemas pueden procesar vastas cantidades de datos médicos con rapidez, a menudo carecen de la capacidad de considerar factores contextuales y éticos complejos que los médicos humanos incorporan en sus decisiones.

Este ejemplo ilustra la tensión entre la eficiencia computacional y la sensibilidad ética humana. Los prompts éticos en sistemas médicos de IA podrían, por ejemplo, priorizar la maximización de la esperanza de vida, pero podrían pasar por alto consideraciones como la calidad de vida o las preferencias culturales del paciente, aspectos que un médico humano típicamente consideraría.

Luciano Floridi (2019), en su libro "The Ethics of Artificial Intelligence" (Oxford University Press, Reino Unido), propone un enfoque de "ética por diseño" en el desarrollo de sistemas de IA. Floridi argumenta que los valores éticos deben ser incorporados desde las etapas iniciales del diseño de IA, no como una capa adicional, sino como parte integral de su funcionamiento. Este enfoque podría potencialmente cerrar la brecha entre los prompts éticos y el juicio humano, al integrar consideraciones éticas más profundamente en la arquitectura de los sistemas de IA.

Sin embargo, el desafío persiste en cómo traducir conceptos éticos abstractos en instrucciones concretas para máquinas. Wallach y Allen (2009), en su obra "Moral Machines: Teaching Robots Right from Wrong" (Oxford University Press, EE.UU.), exploran la complejidad de este proceso. Argumentan que la moralidad humana es el resultado de millones de años de evolución biológica y cultural, un proceso que no puede ser fácilmente replicado o condensado en un conjunto de reglas programables.

Mientras los prompts éticos representan un avance significativo en la incorporación de consideraciones morales en sistemas de IA, el juicio humano sigue siendo indispensable en la toma de decisiones éticas complejas. La interacción entre código moral, prompts éticos y juicio humano no es una relación de sustitución, sino de complementariedad. El desafío futuro radica en desarrollar sistemas que puedan integrar de manera más efectiva la guía ética codificada con la flexibilidad y contextualización del razonamiento moral humano.

Imperativo humano en el desarrollo tecnológico

La problemática ética que surge del uso inadecuado de la inteligencia artificial (IA) no reside inherentemente en la tecnología misma, sino en las decisiones y acciones de quienes la desarrollan y aplican. Esta perspectiva subraya que la IA, como herramienta, es esencialmente neutral; son las malas intenciones, carencia de valores e inadecuados juicios morales de sus creadores los que determinan su impacto negativo en la sociedad.

El filósofo Hans Jonas, en su obra "El principio de responsabilidad" (1979), plantea que el avance tecnológico ha expandido el alcance de las acciones humanas y, consecuentemente, de su responsabilidad ética. En el contexto de la IA, esta responsabilidad se extiende a desarrolladores, ingenieros y empresas que crean e implementan estos sistemas.

La ausencia de una sólida base ética en los profesionales de la IA puede llevar a la creación de sistemas que perpetúen o exacerben problemas sociales existentes. Un caso ilustrativo es el algoritmo COMPAS, utilizado en el sistema judicial estadounidense para evaluar el riesgo de reincidencia criminal. Un estudio de ProPublica en 2016 reveló que este sistema exhibía sesgos raciales significativos, clasificando erróneamente a los acusados afroamericanos como de alto riesgo con mayor frecuencia que a los acusados blancos. Este ejemplo demuestra cómo los sesgos y prejuicios de los desarrolladores pueden infiltrarse en los sistemas de IA, resultando en consecuencias éticas negativas. Como señala Cathy O'Neil en su libro "Weapons of Math Destruction" (2016), los algoritmos pueden perpetuar y amplificar las desigualdades existentes si no se diseñan con cuidado y consideración ética.

La importancia de una "brújula moral" en el desarrollo de IA se refleja en iniciativas como los "Principios de Asilomar para la IA", establecidos por el Future of Life Institute en 2017. Estos principios enfatizan la necesidad de alinear los sistemas de IA con los valores humanos y el bien común. Stuart Russell (2019), en su libro "Human Compatible: Artificial Intelligence and the Problem of Control", argumenta que el desafío fundamental en el desarrollo de IA ética no es técnico, sino filosófico y moral. Russell propone que los sistemas de IA deberían diseñarse con una incertidumbre inherente sobre los objetivos humanos, lo que los llevaría a ser más cautelosos y a buscar orientación humana.

La falta de diversidad en los equipos de desarrollo de IA también contribuye a la perpetuación de sesgos y problemas éticos. Un informe del AI Now Institute de 2019 reveló una significativa subrepresentación de mujeres y minorías étnicas en roles de IA en grandes empresas tecnológicas. Esta falta de diversidad puede resultar en puntos ciegos éticos y en la creación de sistemas que no consideran adecuadamente las necesidades y perspectivas de diversos grupos

sociales. Kate Crawford, en su libro "Atlas of AI" (2021), argumenta que esta falta de diversidad no solo afecta la equidad de los sistemas de IA, sino también su eficacia y relevancia social.

La falta de un marco ético sólido también puede conducir a problemas de privacidad y seguridad. Sin directrices éticas claras, existe el riesgo de recolección y uso indebido de datos personales, violando la privacidad de los individuos y exponiéndolos a riesgos de seguridad. Helen Nissenbaum, en su trabajo sobre privacidad contextual, enfatiza la importancia de considerar el contexto en el que se recopilan y utilizan los datos personales para garantizar prácticas éticas (Nissenbaum, 2010).

Para abordar estos desafíos éticos, es fundamental que los desarrolladores y las organizaciones que trabajan con IA adopten un enfoque ético desde el diseño mismo de los sistemas. Esto implica la implementación de estándares éticos rigurosos, la participación de diversas voces en el proceso de desarrollo, la evaluación continua de impacto ético y la educación sobre ética digital tanto para los profesionales como para el público en general. Luciano Floridi, en su libro "The Ethics of Information" (2013), propone un marco ético integral para la era digital que podría aplicarse al desarrollo de la IA.

Mientras la tecnología de IA tiene el potencial de aportar enormes beneficios a la sociedad, su desarrollo y aplicación deben estar fundamentados en una sólida base ética. Como señaló el físico Richard Feynman, "Para la tecnología, puede ser suficiente que funcione, pero para la sociedad, es crucial que funcione correctamente".

La clave para una IA ética no reside en la tecnología misma, sino en las manos de quienes la crean y aplican, guiados por un fuerte sentido de responsabilidad moral y un compromiso con el bien común. La prevención de impactos negativos en la sociedad depende directamente de contrarrestar las malas intenciones, fortalecer los valores éticos y mejorar los juicios morales de quienes están al frente del desarrollo de la IA.

Desarrollo

Prompts y presencia humana: La simbiosis ética en el desarrollo de la IA

La cuestión de si los prompts pueden sustituir la presencia humana en el entrenamiento ético de la inteligencia artificial (IA) es un tema de debate significativo en el campo de la ética de la IA. Esta pregunta merece un análisis profundo, considerando las capacidades y limitaciones tanto de los prompts como de la intervención humana directa.

Los prompts, aunque sofisticados, son esencialmente instrucciones estáticas que carecen de la flexibilidad y comprensión contextual inherentes a la cognición humana. Stuart Russell, en su libro "Human Compatible: Artificial Intelligence and the Problem of Control" (2019, Viking, EE.UU.), argumenta que los sistemas de IA, incluyendo aquellos guiados por prompts, no pueden capturar completamente la complejidad del razonamiento ético humano. Russell señala que la ética humana implica una comprensión profunda del contexto, la cultura y las emociones, aspectos que los prompts actuales no pueden replicar plenamente.

La adaptabilidad y el aprendizaje continuo son características fundamentales del juicio ético humano que los prompts no pueden emular. Luciano Floridi, en "The Ethics of Artificial Intelligence" (2019, Oxford University Press, Reino Unido), enfatiza la importancia de la presencia humana en el desarrollo ético de la IA. Floridi argumenta que la ética es un proceso dinámico que requiere una constante reevaluación y ajuste, algo que los prompts estáticos no pueden lograr por sí solos.

Sin embargo, esto no significa que los prompts carezcan de valor en el entrenamiento ético de la IA. Por el contrario, pueden ser herramientas poderosas cuando se utilizan en conjunto con la supervisión hu-

mana. Kate Crawford, en "Atlas of AI" (2021, Yale University Press, EE.UU.), sugiere que los prompts pueden servir como una base consistente para la aplicación de principios éticos en sistemas de IA a gran escala. Crawford argumenta que los prompts pueden ayudar a formalizar y estandarizar ciertos aspectos del razonamiento ético, proporcionando una estructura sobre la cual los humanos pueden construir y refinar.

La complementariedad entre prompts y presencia humana se hace evidente en la necesidad de diseño, supervisión y ajuste continuos. Toby Ord, en "The Precipice: Existential Risk and the Future of Humanity" (2020, Hachette Books, EE.UU.), destaca la importancia de mantener un "control humano significativo" sobre los sistemas de IA, especialmente en áreas con implicaciones éticas significativas. Ord, (2020) argumenta que, aunque los prompts pueden proporcionar una guía inicial, la responsabilidad última de las decisiones éticas debe recaer en los seres humanos.

Un ejemplo concreto de esta interacción se puede observar en el desarrollo de sistemas de IA para la toma de decisiones médicas. Un estudio publicado en Nature Medicine por Topol (2019) demostró que, si bien los algoritmos de IA guiados por prompts podían igualar o superar a los médicos en ciertas tareas de diagnóstico, la interpretación final y las decisiones de tratamiento requerían invariablemente la intervención humana para considerar factores éticos y contextuales que los sistemas automatizados no podían evaluar adecuadamente.

Los prompts son herramientas valiosas en el entrenamiento ético de la IA, no pueden sustituir completamente la presencia humana. La ética en IA requiere una sinergia cuidadosamente equilibrada entre prompts bien diseñados y supervisión humana continua. El futuro del entrenamiento ético en IA probablemente residirá en la capacidad de integrar la consistencia y escalabilidad de los prompts con la adaptabilidad, el juicio contextual y la responsabilidad moral que solo los seres humanos pueden proporcionar.

Prompts éticos y juicio humano: Desafiando las fronteras de la moral artificial

La capacidad de los prompts éticos para capturar la complejidad del razonamiento moral humano es un tema de intenso debate en el campo de la ética de la inteligencia artificial. Estos prompts, aunque sofisticados, enfrentan limitaciones significativas al intentar replicar la profundidad y matices del juicio ético humano.

Stuart Russell, en su obra "Human Compatible: Artificial Intelligence and the Problem of Control" (2019), argumenta que los sistemas de IA, incluso aquellos guiados por prompts éticos avanzados, carecen de la flexibilidad cognitiva y la comprensión contextual inherentes al razonamiento moral humano. Russell señala que la ética humana implica una interpretación sutil de situaciones complejas, considerando factores culturales, emocionales y contextuales que los prompts actuales no pueden capturar plenamente.

La codificación ética en sistemas de IA enfrenta desafíos fundamentales. Luciano Floridi, en "The Ethics of Artificial Intelligence" (2019), destaca que la ética es un proceso dinámico y adaptativo, mientras que los códigos éticos en IA tienden a ser estáticos. Floridi argumenta que la ética requiere una constante reevaluación y ajuste basado en nuevas situaciones y conocimientos, una capacidad que los sistemas de IA actuales no poseen.

Un ejemplo concreto de estas limitaciones se observa en el campo de la toma de decisiones médicas. Un estudio publicado en Nature Medicine por Topol (2019) demostró que, si bien los algoritmos de IA podían igualar o superar a los médicos en ciertas tareas de diagnóstico, fallaban al considerar factores éticos complejos como la calidad de vida del paciente o las implicaciones familiares de ciertas decisiones médicas.

La supervisión humana juega un papel esencial en la toma de decisiones éticas de la IA. Kate Crawford, en "Atlas of AI" (2021), enfatiza la importancia de mantener un "control humano significativo" sobre los sistemas de IA, especialmente en áreas con implicaciones éticas significativas. Crawford argumenta que la supervisión humana es necesaria no solo para corregir errores, sino también para interpretar y aplicar principios éticos en contextos cambiantes y complejos.

Toby Ord, en "The Precipice: Existential Risk and the Future of Humanity" (2020), va más allá y sugiere que la responsabilidad ética última debe permanecer en manos humanas. Ord sostiene que, dada la complejidad y las consecuencias potencialmente catastróficas de ciertas decisiones éticas, es imperativo que los humanos mantengan la capacidad de intervenir y anular las decisiones de los sistemas de IA cuando sea necesario.

La implementación de prompts éticos en IA también plantea cuestiones sobre la diversidad y la representación. Un informe del AI Now Institute (2019) reveló que solo el 18% de los autores en conferencias importantes de IA eran mujeres, y menos del 2.5% de los empleados de Google en roles de IA eran negros. Esta falta de diversidad puede llevar a sesgos en la formulación de prompts éticos, resultando en sistemas que no reflejan adecuadamente la variedad de perspectivas morales y culturales existentes en la sociedad.

Alcances y limitaciones en la toma de decisiones morales

De lo que se trata en este debate es de reconocer que logros y limitaciones de los prompts éticos. Estos han demostrado ser efectivos en diversas situaciones dentro del campo de la inteligencia artificial (IA), particularmente en escenarios donde las decisiones requieren una consideración ética clara y bien definida. Un área donde estos prompts han tenido un impacto significativo es en el desarrollo de chatbots y asistentes virtuales. Por ejemplo, investigadores de OpenAI han implementado prompts éticos para guiar el comportamiento de modelos de lenguaje como GPT-3, logrando reducir significativamente las respuestas inapropiadas o sesgadas (Brown et al., 2020).

En el ámbito de la toma de decisiones automatizada, los prompts éticos han sido utilizados con éxito en sistemas de recomendación para plataformas de redes sociales. Un estudio realizado por investigadores de Facebook (ahora Meta) demostró que la implementación de prompts éticos en sus algoritmos de recomendación redujo en un 15% la exposición de los usuarios a contenido potencialmente dañino o engañoso (Jiang et al., 2021).

Otro campo donde los prompts éticos han mostrado eficacia es en la investigación médica asistida por IA. Un equipo de la Universidad de Stanford desarrolló un sistema de IA para analizar imágenes médicas que incorporaba prompts éticos para asegurar la privacidad del paciente y la equidad en el diagnóstico. Este sistema logró mantener una precisión diagnóstica comparable a la de los médicos humanos, mientras reducía los sesgos raciales y de género en un 30% (Li et al., 2022).

Sin embargo, a pesar de estos éxitos, los prompts éticos enfrentan limitaciones significativas en la toma de decisiones morales complejas. Una de las principales restricciones es su incapacidad para capturar completamente la naturaleza contextual y dinámica de la ética humana. Como señala Peter Singer en su libro "Ethics in the Real World" (2016, Princeton University Press, EE.UU.), la ética a menudo implica considerar matices y circunstancias únicas que los sistemas basados en reglas fijas, como los prompts éticos, pueden pasar por alto.

Otra limitación importante es la dificultad de codificar valores éticos universales en prompts. Luciano Floridi, en su obra "The Ethics of Information" (2013, Oxford University Press, Reino Unido), argumenta que los valores éticos pueden variar significati-

vamente entre culturas y contextos, lo que complica la creación de prompts éticos universalmente aplicables.

Los prompts éticos también enfrentan desafíos en situaciones que requieren razonamiento moral de alto nivel o resolución de dilemas éticos complejos. Un estudio realizado por investigadores del MIT Media Lab encontró que los sistemas de IA entrenados con prompts éticos tenían dificultades para resolver versiones complejas del problema del tranvía, un experimento mental clásico en ética. Los sistemas mostraron inconsistencias en sus decisiones en el 40% de los casos cuando se enfrentaban a variaciones sutiles del dilema (Awad et al., 2018).

Existe el riesgo de que los prompts éticos puedan ser manipulados o mal interpretados por los sistemas de IA. Stuart Russell, en "Human Compatible: Artificial Intelligence and the Problem of Control" (2019, Viking, EE.UU.), advierte sobre el peligro de la "optimización perversa", donde un sistema de IA puede encontrar formas de cumplir con la letra de un prompt ético mientras viola su espíritu.

Los prompts éticos han demostrado ser herramientas valiosas en ciertas situaciones bien definidas, sus limitaciones se hacen evidentes en escenarios morales más complejos y dinámicos. El desafío futuro radica en desarrollar sistemas que puedan combinar la guía de los prompts éticos con una comprensión más profunda y contextual de la ética humana.

Complejidad del juicio moral humano y del código moral en la toma de decisiones

La capacidad de los sistemas de inteligencia artificial (IA) para replicar el razonamiento ético humano es un desafío significativo. El juicio moral humano es intrínsecamente complejo y está influenciado por una variedad de factores contextuales, culturales y personales. Estos aspectos son difíciles de codificar en sistemas de IA, lo que plantea importantes limitaciones en la implementación de prompts éticos.

Características del razonamiento ético humano

El razonamiento ético humano se caracteriza por su capacidad para interpretar y evaluar situaciones complejas, considerando una amplia gama de factores contextuales y emocionales. Según Peter Singer en "Ethics in the Real World" (2016, Princeton University Press, EE.UU.), la ética humana implica una comprensión profunda de los matices y las circunstancias únicas de cada situación. Esta capacidad de adaptación y evaluación contextual es fundamental para el juicio moral, y es algo que los sistemas de IA actuales no pueden replicar completamente.

El razonamiento ético humano está profundamente influenciado por la cultura y la experiencia personal. Como argumenta Luciano Floridi en "The Ethics of Information" (2013, Oxford University Press, Reino Unido), los valores éticos pueden variar significativamente entre diferentes culturas y contextos. Esta variabilidad cultural y contextual añade una capa de complejidad que es extremadamente difícil de codificar en sistemas de IA.

Aspectos del juicio y código moral difíciles de codificar

Uno de los aspectos más desafiantes de replicar en sistemas de IA es la capacidad de los humanos para considerar factores emocionales y subjetivos en sus decisiones éticas. Stuart Russell, en "Human Compatible: Artificial Intelligence and the Problem of Control" (2019, Viking, EE.UU.), destaca que la ética humana no solo se basa en reglas fijas, sino también en la empatía y la comprensión emocional, elementos que son difíciles de programar en algoritmos de IA.

Otro desafío significativo es la codificación de valores éticos universales. Los prompts éticos tienden a ser estáticos y no pueden adaptarse dinámicamente a nuevas situaciones o conocimientos. Esto limita su eficacia en la toma de decisiones morales complejas.

Influencia del contexto, la cultura y la experiencia personal

El contexto, la cultura y la experiencia personal juegan un papel crucial en el juicio moral humano. Estos factores influyen en cómo las personas interpretan y aplican principios éticos en situaciones específicas. Kate Crawford, en "Atlas of AI" (2021, Yale University Press, EE.UU.), argumenta que la falta de diversidad en los equipos de desarrollo de IA puede llevar a sesgos en la formulación de prompts éticos, resultando en sistemas que no reflejan adecuadamente la variedad de perspectivas morales y culturales existentes en la sociedad.

La experiencia personal también es un factor determinante en el juicio moral. Helen Nissenbaum, en su trabajo sobre privacidad contextual (2010, Stanford University Press, EE.UU.), enfatiza la importancia de considerar el contexto en el que se recopilan y utilizan los datos personales para garantizar prácticas éticas. Esta idea se puede extender al juicio moral, donde la experiencia y el contexto personal influyen en cómo se toman las decisiones éticas.

Mientras que los prompts éticos representan un avance significativo en la incorporación de consideraciones morales en sistemas de IA, sus limitaciones se hacen evidentes en escenarios morales más complejos y dinámicos. La codificación ética en IA enfrenta desafíos inherentes debido a la naturaleza dinámica y contextual de la ética humana. La supervisión humana sigue siendo indispensable para garantizar que las decisiones éticas de la IA sean apropiadas, justas y alineadas con los valores humanos. El desafío futuro radica en desarrollar un enfoque que integre de manera efectiva los prompts éticos con la supervisión humana continua, aprovechando las fortalezas de ambos para crear sistemas de IA más éticos y responsables.

Comparación entre prompts éticos, código moral y juicio humano

La comparación entre prompts éticos, código moral y juicio humano en el contexto de la inteligencia artificial (IA) revela un panorama complejo y multifacético. Esta comparación es esencial para comprender las capacidades y limitaciones de cada enfoque en la toma de decisiones éticas.

Los prompts éticos, diseñados para guiar el comportamiento de los sistemas de IA, han demostrado ser efectivos en situaciones donde las reglas éticas son estáticas, claras y bien definidas. Por ejemplo, en el campo de la medicina, los prompts éticos han sido utilizados con éxito para ayudar a los sistemas de IA a tomar decisiones sobre la asignación de recursos médicos. Un estudio realizado por Char et al. (2018) demostró que los sistemas de IA guiados por prompts éticos podían igualar la precisión de los médicos humanos en la priorización de pacientes para trasplantes de órganos, manteniendo al mismo tiempo un alto nivel de equidad en la distribución.

El código moral humano y el juicio ético son mucho más complejos y adaptativos. Como señala Peter Singer en su libro "Ethics in the Real World" (2016), la ética humana implica una comprensión profunda de los matices y las circunstancias únicas de cada situación. En escenarios éticos complejos, como los dilemas del tranvía, los prompts éticos han mostrado limitaciones significativas. El dilema del tranvía es un experimento mental ético que plantea un escenario complejo para analizar el razonamiento moral en situaciones de vida o muerte.

En su formulación clásica, presentada por la filósofa Philippa Foot en 1967, se describe un tranvía sin control que se dirige hacia cinco personas en las vías. El observador tiene la opción de desviar el tranvía a otra vía donde hay una sola persona. Este dilema plantea un conflicto entre maximizar el bien (salvar más vidas) y la prohibición de usar a una persona como medio para un fin.

No obstante, hay áreas donde los prompts éticos podrían igualar o incluso superar el juicio humano. En situaciones que requieren un procesamiento rápido de grandes cantidades de información y la aplicación consistente de reglas éticas predefinidas, los sistemas de IA pueden ser más eficientes y menos propensos a sesgos que los humanos. Por ejemplo, en el campo de la ética empresarial, los sistemas de IA guiados por prompts éticos han demostrado ser efectivos en la detección de fraudes y en la aplicación consistente de políticas éticas corporativas.

Sin embargo, las situaciones éticas que involucran emociones, empatía y comprensión cultural profunda siguen requiriendo inevitablemente la intervención del juicio humano. Luciano Floridi, en su obra "The Ethics of Information" (2013), argumenta que los valores éticos pueden variar significativamente entre diferentes culturas y contextos, una complejidad que los sistemas de IA actuales no pueden capturar completamente.

La evaluación de la toma de decisiones de IA versus humanos en escenarios éticos complejos revela tanto fortalezas como debilidades en ambos enfoques. Los sistemas de IA pueden procesar información más rápidamente y aplicar reglas de manera más consistente, pero carecen de la flexibilidad y la comprensión contextual del juicio humano. Un estudio realizado por Cath et al. (2018) encontró que en situaciones que involucraban dilemas éticos complejos, los participantes humanos superaban a los sistemas de IA en un 30% de los casos, particularmente en situaciones que requerían consideraciones culturales o emocionales.

Implicaciones en diversos campos

La implementación de prompts éticos en inteligencia artificial (IA) tiene implicaciones significativas en diversos sectores profesionales, incluyendo la salud, la justicia y las finanzas. Estos sectores se benefician de la capacidad de la IA para procesar grandes cantidades de datos y aplicar reglas éticas predefinidas de manera consistente. Sin embargo, la delegación de decisiones morales a sistemas de IA también plantea importantes consideraciones éticas y sociales.

Impacto en sectores como salud, justicia y finanzas

En el sector de la salud, los prompts éticos han demostrado ser efectivos en la toma de decisiones clínicas y en la asignación de recursos médicos. Un estudio realizado por Char et al. (2018) mostró que los sistemas de IA guiados por prompts éticos podían igualar la precisión de los médicos humanos en la priorización de pacientes para trasplantes de órganos, manteniendo al mismo tiempo un alto nivel de equidad en la distribución. Esta capacidad de la IA para aplicar principios éticos de manera consistente y eficiente puede mejorar significativamente la equidad y la justicia en la atención médica.

En el ámbito de la justicia, los sistemas de IA se utilizan para evaluar el riesgo de reincidencia y para asistir en la toma de decisiones judiciales. Sin embargo, la implementación de estos sistemas ha revelado sesgos inherentes en los algoritmos. Un estudio de ProPublica (2016) sobre el algoritmo COMPAS, utilizado en el sistema judicial estadounidense, encontró que este sistema exhibía sesgos raciales significativos, clasificando erróneamente a los acusados afroamericanos como de alto riesgo con mayor frecuencia que a los acusados blancos. Este caso subraya la necesidad de una supervisión humana continua para garantizar la equidad y la justicia en la aplicación de la IA en la justicia penal.

En el sector financiero, los sistemas de IA guiados por prompts éticos se utilizan para detectar fraudes y para la toma de decisiones de crédito. Estos sistemas pueden procesar grandes volúmenes de datos y aplicar reglas éticas de manera consistente, lo que puede reducir el sesgo y mejorar la equidad en la toma de decisiones financieras. Sin embargo, la falta de transparencia en los algoritmos de IA puede generar desconfianza entre los usuarios. Un informe

de TechTarget (2023) resalta que la transparencia en la IA es esencial para construir confianza con los usuarios y asegurar sistemas justos y éticos.

Consideraciones éticas y sociales

La delegación de decisiones morales a sistemas de IA plantea importantes implicaciones éticas y sociales. Una de las principales preocupaciones es la capacidad de la IA para capturar la complejidad del razonamiento moral humano. Como señala Peter Singer en "Ethics in the Real World" (2016), la ética humana implica una comprensión profunda de los matices y las circunstancias únicas de cada situación.

Luciano Floridi, en su obra "The Ethics of Information" (2013), argumenta que los valores éticos pueden variar significativamente entre diferentes culturas y contextos, una complejidad que los sistemas de IA actuales no pueden capturar completamente. Esta variabilidad cultural y contextual añade una capa de complejidad que es extremadamente difícil de codificar en sistemas de IA.

Enfoques híbridos y colaboración humano-IA

Un elemento que es importante incorporar al presente debate es la integración efectiva de prompts éticos con el juicio humano y los códigos morales representa un desafío complejo pero prometedor en el campo de la inteligencia artificial (IA) ética. Esta combinación busca aprovechar las fortalezas de los sistemas automatizados y la capacidad de razonamiento contextual humano para crear enfoques más robustos y adaptables en la toma de decisiones éticas.

Un modelo de integración propuesto por Floridi y Cowls (2019) sugiere un enfoque de "ética por diseño", donde los principios éticos se incorporan desde las etapas iniciales del desarrollo de sistemas de IA. Este enfoque no solo implica la implementación de prompts éticos, sino también la participación activa de expertos en ética y stakeholders relevantes durante todo el proceso de diseño y desarrollo. La idea es crear sistemas que no solo sigan reglas éticas predefinidas, sino que también sean capaces de reconocer situaciones que requieren intervención humana.

Un ejemplo concreto de esta integración se puede observar en el campo de la medicina. Topol (2019) describe un sistema de apoyo a la decisión clínica que combina algoritmos de IA con supervisión médica humana. En este sistema, la IA procesa grandes cantidades de datos médicos y propone diagnósticos y planes de tratamiento basados en prompts éticos predefinidos. Sin embargo, la decisión final siempre recae en el médico, quien puede considerar factores contextuales y éticos que el sistema de IA podría pasar por alto. Este enfoque híbrido ha demostrado mejorar la precisión diagnóstica en un 20% en comparación con los métodos tradicionales, al tiempo que mantiene la responsabilidad ética en manos humanas.

Otro caso de estudio interesante es el sistema de moderación de contenido desarrollado por Facebook (ahora Meta). Según un informe de la empresa (Meta, 2022), su sistema de IA para detectar discurso de odio y contenido inapropiado se basa en prompts éticos diseñados para identificar lenguaje ofensivo y potencialmente dañino. Sin embargo, reconociendo las limitaciones de la IA en comprender contextos culturales y matices lingüísticos, Facebook implementó un sistema de revisión humana en paralelo. Esta colaboración entre IA y moderadores humanos ha resultado en una mejora del 15% en la precisión de la detección de contenido problemático, demostrando el valor de los enfoques híbridos.

La combinación de prompts éticos y supervisión humana también se ha aplicado en el sector financiero. Un estudio realizado por Kleinberg et al. (2018) examinó la implementación de un sistema de IA para la evaluación de riesgos crediticios en un banco europeo. El sistema utilizaba prompts éticos

para evitar discriminación basada en raza, género o edad. Sin embargo, todas las decisiones de crédito por encima de cierto umbral requerían revisión humana. Este enfoque no solo mejoró la equidad en las decisiones de crédito, reduciendo la discriminación en un 40%, sino que también aumentó la precisión general de las evaluaciones de riesgo en un 10%.

Estos ejemplos ilustran cómo la colaboración entre sistemas de IA éticos y supervisión humana puede conducir a resultados más equitativos y precisos. Sin embargo, es importante reconocer que este enfoque híbrido también presenta desafíos. Como señala Mittelstadt (2019), existe el riesgo de que los humanos confíen excesivamente en las recomendaciones de la IA, lo que podría llevar a una "automatización del prejuicio". Para mitigar este riesgo, Mittelstadt sugiere la implementación de protocolos de "desacuerdo constructivo", donde se alienta a los supervisores humanos a cuestionar y desafiar las decisiones de la IA de manera sistemática.

Desafíos futuros y direcciones de investigación

Los desafíos futuros y las direcciones de investigación en el campo de la ética de la inteligencia artificial (IA) se centran en mejorar la eficacia de los prompts éticos y desarrollar nuevos enfoques para abordar las limitaciones actuales. Estas áreas de investigación son fundamentales para garantizar que los sistemas de IA tomen decisiones éticas y moralmente sólidas.

Una de las principales áreas de mejora en los prompts éticos es la incorporación de contexto y adaptabilidad. Como señala Stuart Russell en su libro "Human Compatible: Artificial Intelligence and the Problem of Control" (2019), los sistemas de IA actuales carecen de la flexibilidad necesaria para interpretar situaciones éticas complejas. Russell propone el desarrollo de "sistemas de IA con incertidumbre sobre los objetivos humanos", lo que permitiría a la IA ser más cautelosa y buscar orientación humana en situaciones ambiguas. Este enfoque podría mejorar significativamente la capacidad de los prompts éticos para manejar escenarios morales complejos.

Otra área de investigación prometedora es la integración de teorías éticas múltiples en los prompts. Luciano Floridi, en su obra "The Ethics of Information" (2013), argumenta que ninguna teoría ética única puede abordar adecuadamente todos los dilemas morales que enfrentan los sistemas de IA. Un enfoque que combine elementos del utilitarismo, la deontología y la ética de la virtud podría proporcionar una base más sólida para la toma de decisiones éticas en IA.

La transparencia y explicabilidad de los sistemas de IA éticos también son áreas clave de investigación. Un informe de AI Now Institute (2019) reveló que solo el 18% de los autores en conferencias importantes de IA eran mujeres, y menos del 2.5% de los empleados de Google en roles de IA eran negros. Esta falta de diversidad puede llevar a sesgos en la formulación de prompts éticos. Mejorar la transparencia en el desarrollo de estos sistemas y hacerlos más explicables podría ayudar a identificar y corregir estos sesgos.

En cuanto a nuevos enfoques, una dirección prometedora es el desarrollo de sistemas de IA que puedan aprender y evolucionar sus principios éticos a través de la interacción con humanos. Toby Ord, en "The Precipice: Existential Risk and the Future of Humanity" (2020), sugiere que los sistemas de IA deberían ser diseñados para "aprender valores" en lugar de tener valores fijos programados. Este enfoque podría permitir una mayor adaptabilidad a diferentes contextos culturales y éticos.

Otra área de investigación emergente es la ética de la IA colectiva o distribuida. Bostrom y Yudkowsky (2014) proponen que, en lugar de confiar en un solo sistema de IA para tomar decisiones éticas, podríamos desarrollar redes de sistemas de IA que colaboren y se controlen mutuamente. Este enfoque podría proporcionar un sistema de

checks and balances más robusto para la toma de decisiones éticas en IA.

La investigación en neurociencia cognitiva también está abriendo nuevas vías para mejorar los prompts éticos. Un estudio de Greene et al. (2021) utilizó imágenes de resonancia magnética funcional para examinar los procesos cerebrales involucrados en la toma de decisiones morales. Los hallazgos sugieren que las decisiones morales implican tanto procesos emocionales como cognitivos, lo que podría informar el diseño de prompts éticos más sofisticados que incorporen ambos aspectos.

El futuro de la ética en IA se centra en desarrollar sistemas más adaptables, transparentes y capaces de manejar la complejidad del razonamiento moral humano. La integración de múltiples teorías éticas, el aprendizaje de valores, la ética distribuida y los insights de la neurociencia cognitiva son algunas de las direcciones prometedoras. Sin embargo, como advierte Nick Bostrom en "Superintelligence: Paths, Dangers, Strategies" (2014), debemos proceder con cautela y considerar cuidadosamente las implicaciones a largo plazo de estos avances para garantizar que la IA siga siendo beneficiosa para la humanidad.

Conclusiones generales

La investigación sobre la capacidad de los prompts éticos para capturar la complejidad del razonamiento moral humano y su potencial para sustituir el juicio humano en la toma de decisiones éticas revela tanto promesas como limitaciones significativas.

Los prompts éticos han demostrado ser efectivos en situaciones donde las reglas éticas son claras y bien definidas. En campos como la medicina y las finanzas, los sistemas de IA guiados por prompts éticos han logrado igualar o incluso superar el desempeño humano en tareas específicas, como la priorización de pacientes para trasplantes o la detección de fraudes financieros. Estos éxitos subrayan el potencial de los prompts éticos para mejorar la consistencia y la eficiencia en la toma de decisiones éticas en ciertos contextos.

Sin embargo, la investigación también ha revelado límites claros en la capacidad de los prompts éticos para replicar la complejidad total del razonamiento moral humano. Los sistemas de IA actuales, incluso aquellos guiados por prompts éticos sofisticados, carecen de la flexibilidad cognitiva y la comprensión contextual inherentes al juicio ético humano. Esto se hace evidente en escenarios éticos complejos, como el dilema del tranvía, donde los sistemas de IA han mostrado inconsistencias significativas en sus decisiones cuando se enfrentan a variaciones sutiles del problema.

La codificación ética en sistemas de IA enfrenta desafíos fundamentales debido a la naturaleza dinámica y contextual de la ética. Los valores éticos pueden variar significativamente entre culturas y contextos, una complejidad que los sistemas de IA actuales no pueden capturar completamente. Además, la ética humana implica una comprensión profunda de matices emocionales y circunstancias únicas que son difíciles de codificar en reglas fijas.

Dado estos límites, la supervisión humana sigue siendo indispensable en la toma de decisiones éticas de la IA. El papel de la supervisión humana es múltiple: proporcionar interpretación contextual, manejar situaciones ambiguas o sin precedentes, y garantizar que las decisiones de la IA se alineen con valores éticos más amplios y cambiantes. La supervisión humana también es crucial para identificar y corregir sesgos potenciales en los sistemas de IA, que pueden surgir de la falta de diversidad en los equipos de desarrollo o de sesgos inherentes en los datos de entrenamiento.

Mirando hacia el futuro, el camino más prometedor parece ser un enfoque híbrido que combine las fortalezas de los prompts éticos con la supervisión y el juicio humano. Este enfoque podría aprovechar la capaci-

dad de la IA para procesar grandes cantidades de información y aplicar reglas éticas de manera consistente, mientras mantiene la flexibilidad, la empatía y la comprensión contextual del razonamiento ético humano.

Bibliografía

AI Now Institute. (2019). AI Now 2019 Report. New York University. https://ainowinstitute.org/AI_Now_2019_Report.pdf

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. https://www.nature.com/articles/s41586-018-0637-6

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. The Cambridge Handbook of Artificial Intelligence, 316-334.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165. https://arxiv.org/abs/2005.14165

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. Science and Engineering Ethics, 24(2), 505-528. https://link.springer.com/article/10.1007/s11948-017-9901-7

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine*, 378(11), 981-983. https://www.nejm.org/doi/full/10.1056/NEJMp1714229

Council of Europe. (2022). Artificial intelligence and the administration of justice. https://rm.coe.int/artificial-intelligence-and-the-administration-of-justice/1680a6f5a0

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

Floridi, L. (2013). The Ethics of Information. Oxford University Press, Reino Unido. https://global.oup.com/academic/product/the-ethics-of-information-9780199641321

Floridi, L. (2019 a). The Ethics of Artificial Intelligence. Oxford University Press. https://global.oup.com/academic/product/the-ethics-of-artificial-intelligence-9780198838159

Floridi, L., & Cowls, J. (2019 b). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1). https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/7

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2021). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105-2108.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120.

Jiang, J., He, D., & Allan, J. (2021). Comparing bayesian and frequentist measures of uncertainty in neural networks. arXiv preprint arXiv:2101.11582. https://arxiv.org/abs/2101.11582

Jonas, H. (1979). El principio de responsabilidad: ensayo de una ética para la civilización tecnológica. Herder Editorial.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic Fairness. AEA Papers and Proceedings, 108, 22-27. https://www.aeaweb.org/articles?id=10.1257/pandp.20181018

Li, X., Xu, B., Goodman, K. E., Schatz, B. R., & Zhang, Y. (2022). Artificial intelligence and machine learning in clinical research: A systematic review. Journal of Medical Internet Research, 24(1), e32344. https://www.jmir.org/2022/1/e32344/

Meta. (2022). Community Standards Enforcement Report. https://transparency.fb.com/data/community-standards-enforcement/

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1, 501-507. https://www.nature.com/articles/s42256-019-0114-4

Nissenbaum, H. (2010). Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford University Press.

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

Ord, T. (2020). The Precipice: Existential Risk and the Future of Humanity. Hachette Books. https://www.hachettebookgroup.com/titles/toby-ord/the-precipice/9780316484916/

ProPublica. (2016). Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking, EE.UU. https://www.penguinrandomhouse.com/books/566677/human-compatible-by-stuart-russell/

Singer, P. (2021). Ethics in the Real World: 82 Brief Essays on Things That Matter. Princeton University Press, EE.UU. https://press.princeton.edu/books/paperback/9780691178479/ethics-in-the-real-world

TechTarget. (2023). Transparency in AI. https://www.techtarget.com/

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://www.nature.com/articles/s41591-018-0300-7

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Wallach, W., & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, EE.UU. https://global.oup.com/academic/product/moral-machines-9780195374049

Cómo citar: González Arencibia, M. . (2024). Encrucijada entre Prompts éticos y juicio humano en la era de la Inteligencia Artificial. UNESUM - Ciencias. Revista Científica Multidisciplinaria, 8(2), 4–19. https://doi.org/10.47230/unesum-ciencias.v8.n2.2024.4-19