RESONANCIAS. Revista de Filosofía ▪ N° 17 ▪ JULIO 2024: 101-120 ▪ ISSN: 0719-790X

Una mirada crítica a la ética de la IA: de preocupaciones emergentes y principios orientadores a un desvelar ético

A Critical Look at AI Ethics: From Emerging Concerns and Guiding Principles to Ethical Unveiling

GABRIELA ARRIAGADA-BRUNEAU[1]
Pontificia Universidad Católica de Chile, Santiago, Chile
gcarriagada@uc.cl
Fecha de recepción: 19/04/2024
Fecha de aceptación: 22/06/2024

Resumen

En este artículo examino el estado actual de la "segunda ola" de la ética en la inteligencia artificial (IA), la cual se centra en la integración de principios éticos fundamentales como la justicia, la privacidad, la transparencia y la explicabilidad en el diseño, uso e implementación de sistemas de IA. Argumento que, aunque esta fase ha sido criticada por su naturaleza abstracta y su falta de contextualización, es imperativo que la emergente "tercera ola" adopte un cambio paradigmático hacia un "desvelar ético". Propongo que este desvelar ético debería involucrar un profundo y continuo proceso hermenéutico que no solo interprete cómo las tecnologías de IA reconfiguran nuestras estructuras sociales, políticas y personales, sino que también actúe como un medio de liberación del enmarcamiento tecnológico Heideggeriano. Este enfoque sugiere que la ética debe ser considerada no solo como un complemento, sino

CÓMO CITAR ESTE ARTÍCULO:

En APA: Arriagada-Bruneau, G. (2024). Una mirada crítica a la ética de la IA: de preocupaciones emergentes y principios orientadores a un desvelar ético. *Resonancias*, (17), 101-120. DOI: 10.5354/0719-790X.2024.74438

En MLA: Arriagada-Bruneau, G. "Una mirada crítica a la ética de la IA: de preocupaciones emergentes y principios orientadores a un desvelar ético." *Resonancias*, n.° 17, julio de 2024, pp. 101-120. DOI: 10.5354/0719-790X.2024.74438

Palabras clave: ética de inteligencia artificial, sistemas sociotécnicos, enmarcamiento tecnológico, principios éticos, ética aplicada

Keywords: Ethics of Artificial Intelligence, Sociotechnical Systems, Technological Framing, Ethical Principles, Applied Ethics

[1] Profesora Asistente, Instituto de Éticas Aplicadas e Instituto de Ingeniería Matemática Computacional, Pontificia Universidad Católica de Chile. https://orcid.org/0000-0002-0006-7024

como un componente esencial y fundamental en el ciclo de vida del desarrollo de la IA, fomentando así una integración más profunda de consideraciones éticas que guíen tanto la innovación tecnológica como su implementación práctica.

Abstract

In this article, I examine the current state of the "second wave" of ethics in artificial intelligence (AI), which focuses on the integration of fundamental ethical principles such as justice, privacy, transparency, and explicability in the design, use, and implementation of AI systems. I argue that although this phase has been criticized for its abstract nature and lack of contextualization, it is imperative that the emerging "third wave" adopts a paradigmatic change towards an "ethical unveiling." I propose that this ethical unveiling should involve a deep and ongoing hermeneutic process that not only interprets how AI technologies reconfigure our social, political, and personal structures but also acts as a means of liberation from the technological framing in the Heideggerian sense. This approach suggests that ethics should not only be considered as an add-on but as an essential and fundamental component in the lifecycle of AI development, thus promoting a deeper integration of ethical considerations that guide both technological innovation and its practical implementation.

1. La segunda ola de la ética de IA: preocupaciones emergentes y principios orientadores

La ética en inteligencia artificial se ha consolidado recientemente como una disciplina. En los últimos años, los artículos de investigación dedicados a esta temática han proliferado. Como notan Borenstein *et al.*, los trabajos académicos en Google Scholar de artículos relacionados a búsquedas sobre "ética de inteligencia artificial" van de 21 resultados en 2016 a 341 en 2011. Es más, recientemente actualizamos la búsqueda con los mismos criterios y los resultados superan los 1000 artículos (referencia eliminada para revisión) y, en general, refieren a discusiones centradas en principios éticos aplicados a la ética de la IA. Esta etapa en el desarrollo de la ética de inteligencia artificial ha sido denominada por Aimee Van Wynsberghe como la segunda ola de la ética de IA. La autora describe la primera ola desde el enfoque en preocupaciones de la superinteligencia y escenarios de riesgo existencial asociado a la "sublevación" de las máquinas, hacienda referencias a uno de los primeros y más reconocidos análisis por Nick Bostrom y Eliezer Yudkowsky.

La segunda ola, según Van Wynsberghe, se ha enfocado en abordar las preocupaciones prácticas relacionadas con el uso de técnicas de aprendizaje automático, los desafíos de transparencia y explicabilidad en las cajas negras algorítmicas, la falta de representatividad en los datos de entrenamiento que

ocasionan sesgos en los modelos de IA, también la violación de la privacidad de los ciudadanos a través de sistemas de reconocimiento facial. En términos simples, esta segunda ola discute la aplicación de principios éticos como la justicia, la privacidad, la transparencia y la explicabilidad en el diseño, uso e implementación de los sistemas de IA.

Ahora bien, para visualizar críticamente esta segunda ola, es relevante considerar preocupaciones emergentes relacionadas con cambios en los contextos de toma de decisiones que, necesariamente, influyen en la adopción de sistemas de IA. Me refiero, por ejemplo, al creciente uso de algoritmos (conjunto finito de instrucciones específicas) para la toma de decisiones. Si bien, no todo algoritmo se utiliza como modelo para entrenar un sistema de IA, el estudio ético del uso de algoritmos como mediadores en procesos sociales nos ha informado mucho sobre preocupaciones emergentes que también se presentan en sistemas de IA.[2]

Recordemos el artículo fundacional relacionado con el uso ético de algoritmos titulado "La ética de los algoritmos: delineando el debate" (Mittelstadt *et al.*), donde se discutió formalmente la creciente influencia de los algoritmos en procesos de decisión que antes dependían únicamente de seres humanos. Ahí, los autores enfatizaban la importancia ética de la mediación algorítmica en procesos sociales, empresariales, gubernamentales, incluso en interacciones personales. Para analizar este fenómeno, reconocieron seis tipos de preocupaciones éticas que se originan en cuestionamientos epistémicos (figura 1). Por ejemplo, la evidencia inconclusa (*inconclusive evidence*) refleja la necesidad de reconocer y gestionar la incertidumbre en sus conclusiones. Asimismo, la evidencia inescrutable (*inscrutable evidence*) establece la importancia de la transparencia y la comprensión del proceso algorítmico que lleva a las conclusiones entregadas por el modelo. Esto se contextualiza en las demandas de una sociedad informada donde la rendición de cuentas (transparencia) y la comprensión pública son pilares para generar confianza en los sistemas tecnológicos. Finalmente, la evidencia equivocada (*misguided evidence*) nos obliga a reflexionar sobre la integridad de los datos de entrada, dado que los errores en las bases de datos pueden llevar a conclusiones erróneas y decisiones injustas.

---

[2] Aunque no hay una definición general de inteligencia artificial que sea aceptada en este trabajo cuando me refiera a sistemas de inteligencia artificial (IA), tendré en mente la definición establecida por el AI HLEG (Grupo de expertos de alto nivel sobre inteligencia artificial) de la Unión Europea. Esta definición dice que la IA es "sistemas que muestran comportamiento inteligente al analizar su entorno y tomar acciones —con cierto grado de autonomía— para alcanzar objetivos específicos". Esta definición, por tanto, se refiere únicamente a lo que se denomina IA estrecha o débil; a saber, que no tiene un espectro completo de habilidades cognitivas en forma de una inteligencia artificial general.

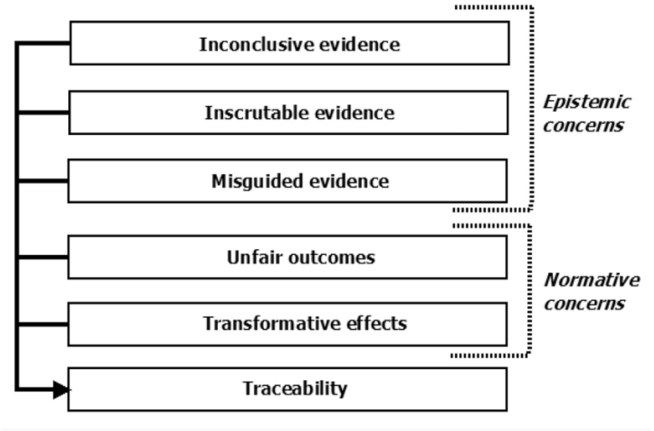

Figure 1. Six types of ethical concerns raised by algorithms.

Figura 1: Seis tipos de preocupaciones éticas sobre uso de algoritmos como mediadores para la toma de decisiones (Mittelstadt *et al.*, 4).

Las siguientes tres preocupaciones se refieren más bien a cuestiones normativas, como los resultados injustos (*unfair outcomes*), donde el enfoque ético requiere una evaluación continua de las acciones que los algoritmos motivan, particularmente cuando podrían reforzar desigualdades existentes o introducir nuevas formas de discriminación. Esto pone de manifiesto la necesidad de criterios éticos robustos que puedan aplicarse al diseño y operación de sistemas algorítmicos para prevenir tales injusticias. En cuanto a los efectos transformativos (*transformative effects*), los algoritmos pueden alterar nuestra percepción del mundo y nuestras estructuras sociales y políticas de maneras significativas y, a menudo, imperceptibles. Por tanto, debemos reflexionar sobre cómo estos cambios influyen en nuestra realidad y los valores que sustentan nuestras sociedades. Finalmente, la trazabilidad (*traceability*) es crucial para poder atribuir responsabilidades cuando se utilizan algoritmos.

Si observamos detenidamente estas categorías de preocupaciones normativas y epistémicas sobre el uso ético de algoritmos, es posible identificar que estas no solo señalaron la urgencia de un enfoque interdisciplinario para abordar los desafíos éticos relacionados con el creciente uso de algoritmos, en particular en contextos de decisión de alto impacto social, sino que también configuran varios de los grandes desarrollos que hemos visto en el avance de la ética de la IA como disciplina. Así, podemos reconocer cómo los avances en esta área se benefician de debates y cuestionamientos en áreas colindantes, como la ética de algoritmos. Por ejemplo, reconocer la incertidumbre inherente en el conocimiento producido por los algoritmos se puede relacionar con la percepción de la tecnología y la IA

como falible (escepticismo epistémico). La evidencia inconclusa generada por los algoritmos nos desafía a ver los avances e incorporaciones de la tecnología como una práctica contextualizada, desde la cual se consideren los valores y prejuicios subyacentes que podrían influir en los resultados de los algoritmos y otras tecnologías, como los sistemas de IA.

En esta línea, autoras como Meredith Broussard subrayan que la IA no es una solución universal ni objetiva y que, por tanto, no debemos caer en un "tecnochovinismo", el cual se define como la creencia de que "la tecnología es siempre la solución" (Broussard, 4). Esta caracterización retrata un optimismo ciego que incita a asumir que la tecnología es menos sesgada que una decisión humana, influenciando no solo las expectativas sobre el funcionamiento de la IA, sino también la evaluación ética que hacemos de esta. Si a esta advertencia le sumamos el factor de "escrutabilidad" y el concepto de "evidencia equivocada" del artículo fundacional sobre la ética de algoritmos, podemos ver cómo en la ética de IA también se nos interpela a cuestionar la validez de los inputs (entradas) y outputs (salidas) de un sistema.

Por esto, una de las preocupaciones epistémico-éticas fundamentales del desarrollo de la inteligencia artificial se ha transformado en discusiones sobre transparencia y explicabilidad respecto a la opacidad y falta de conocimiento sobre qué hace la IA, así como los sesgos y patrones de discriminación que provienen de datos de entrenamiento que reflejan una sociedad altamente sesgada e injusta. Estas son las mismas características que Van Wynsberghe destaca como parte de la segunda ola de la ética de la IA. Como consecuencia, las preocupaciones normativas planteadas por Mittelstadt *et al.* al inicio del debate sobre la ética de algoritmos, lograron permear no solo en directrices para el uso de algoritmos, sino que también se extendieron a la IA, en parte, debido al gran uso de algoritmos como base para modelar sistemas de IA. Así, a medida que la discusión comenzó a proliferar, el mundo académico,[3] que incipientemente se acercaba a estos nuevos desafíos, y el desarrollo tecnológico aplicado a la sociedad, que avanzaba sin marcos, directrices ni estándares que generaran alertas sobre qué se estaba haciendo y cómo, se encontraron en una tormenta perfecta. Este escenario propició el surgimiento de una serie de principios de ética de la IA como respuesta para enfrentar estos avances. No obstante, este pilar fundamental de la segunda ola de la ética de la IA no tardaría en ser criticado por su "inutilidad".

[3] Véase Binns; Boddington; Burrell; Chouldechova; Coeckelbergh; Danks and London; Dencik *et al.*; Dennis *et al.*; Dignum; Etzioni and Etzioni; Floridi *et al.*; Narayanan; 'Robotics'.

1.1 La (in)utilidad de los principios éticos de IA

Uno de los artículos críticos más comentados sobre la proliferación de principios éticos en IA es el de Brent Mittelstadt, titulado "Los principios por sí solos no pueden garantizar una IA ética". Aquí, Mittelstadt señala que muchas de las iniciativas éticas desarrolladas hasta esa fecha se basaban en el uso de estos principios de IA desarrollados durante la segunda ola. Su uso era predominante en la industria, donde, sin embargo, se traducían en un simple caso de "señalización de virtud" (*virtue signaling*) (Mittelstadt 501), en el cual las prácticas éticas se interpretan como tácticas para posponer la regulación (o simular una autorregulación), cambiando el foco de las discusiones a problemas teóricos y respuestas tecnológicas, lo que genera una apariencia de consciencia y enfoque ético. En la práctica, nos dice Mittelstadt, esto implica que los principios éticos de IA son ambiguos y generales, ya que no abordan las discusiones normativas y políticas que subyacen en conceptos clave como la justicia y la privacidad, y pueden, por lo tanto, continuar alimentando un desarrollo de la IA que no profesionaliza buenas prácticas ni logra integrar la ética en sus procesos productivos.

En respuesta a este contexto, Mittelstadt compara la incipiente ética en inteligencia artificial con la ética médica, identificando que hay principios base comunes que podrían estimular un principialismo robusto en IA. Con el respaldo de la OCDE y del Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (AI HLEG), se propusieron cuatro principios para guiar el desarrollo de una IA confiable: respeto por la autonomía humana, prevención de daños, equidad y explicabilidad (High-Level Expert Group on AI). Sin embargo, como bien analiza Mittelstadt, el principialismo de la ética médica surgió en un contexto profesional diametralmente diferente. La ética médica se desarrolla en contextos con una institucionalización y responsabilidades morales profesionales arraigadas en el quehacer médico, con comités de ética e investigación, instituciones médicas fiscalizadoras, cuerpos profesionales y la base transversal que es el juramento hipocrático. Esa robustez permitió que los principios base de la bioética gestaran un lenguaje común en la ética médica (Beauchamp and DeGrazia; Bosk) y que permitiera identificar y conceptualizar los diversos desafíos éticos. El contexto de creación e implementación de la IA es diferente.

Mittelstadt identifica cuatro grandes aspectos que generan las condiciones necesarias para que el principialismo funcione y que, en comparación con la medicina, el desarrollo de la IA carece (502-508): (1) objetivos comunes y deberes fiduciarios, (2) una historia y normas profesionales consolidadas, (3) métodos probados para traducir principios en prácticas y (4) mecanismos robustos de rendición de cuentas legales y profesionales. En comparación con la medicina, que está orientada por un objetivo unificado de promover el bienestar del paciente y se apoya en una estructura profesional bien establecida, normativas claras y

mecanismos de responsabilidad sólidos, beneficiándose de relaciones fiduciarias y un marco regulatorio riguroso que fomenta la confianza y la colaboración. El sector de la IA se enfrenta a un entorno en el cual los intereses de desarrolladores, usuarios y afectados no siempre están alineados, lo que complica la toma de decisiones éticas en un contexto más competitivo que colaborativo (y altamente privatizado), exacerbado por una regulación que varía regionalmente sin establecer deberes fiduciarios claros hacia los sujetos de datos y los usuarios.

En la misma línea, un año después, Thilo Hagendorff criticó las directrices éticas desarrolladas hasta ese entonces, cuestionando que estas no tienen un impacto real transversal en la toma de decisiones humanas en el campo de la IA. Su crítica es clara: "la ética carece de un mecanismo de refuerzo. Las desviaciones de los diversos códigos de ética no tienen consecuencias" (Hagendorff 113). Analizando 22 de las principales directrices de ética de la IA publicadas hasta 2020, Hagendorff argumenta que estas carecen de mecanismos para reforzar sus propias afirmaciones normativas. Así, la ética sirve más para calmar voces críticas del público que para robustecer prácticas en la disciplina.

En relación a este punto, aunque considero que Hagendorff puede tener un argumento sólido sobre cómo estos principios se traducen en prácticas robustas adoptadas por instituciones y profesionales, difiero en su diagnóstico crítico hacia la ética por no tener mecanismos de refuerzo (*reinforcement mechanisms*). El propósito fundamental de la ética no es ejercer coerción o imponer sanciones (en este aspecto nos podemos valer de la regulación), sino más bien guiar el comportamiento humano a través de una reflexión moral y el fomento de la integridad personal y colectiva. Para esto, los principios pueden servir como traductores para definir ciertas preocupaciones éticas dentro del contexto de la IA. La tarea pendiente, no obstante, es integrar la ética como una brújula moral para la acción, que nos permita justificar las decisiones y acciones que tomamos respecto al diseño, uso e implementación de sistemas de IA.

A pesar de esta distinción, comparto el punto crítico de Hagendorff: debido a que las desviaciones de los códigos éticos establecidos no conllevan consecuencias inmediatas, la ética de la IA, a menudo, es percibida como algo externo o adicional a las preocupaciones técnicas, como un marco no vinculante impuesto por instituciones externas a la comunidad técnica. Esta falta de responsabilidad vinculante que apele a códigos y estándares profesionales provoca que, en palabras de Hagendorff, los desarrolladores carezcan de una sensación de responsabilidad o de una percepción de la importancia moral de su trabajo, siendo los incentivos económicos un factor que frecuentemente sobrepasa el compromiso con los principios y valores éticos. Me parece, por tanto, que lo planteado por Hagendorff más bien enfatiza que la ética de la IA necesita lograr integrar estos

principios como parte de las metodologías de trabajo de los desarrolladores de IA y de las instituciones que los financian.

Hoy por hoy, aunque estas críticas han influido en los debates académicos, no hemos logrado superar estas dificultades. Un artículo más reciente de Luke Munn califica nuevamente a la ética de la IA como "inútil", argumentando que la ineficacia de los principios éticos se debe, en gran parte, a que son abstractos, aislados (descontextualizados) y carecen de fuerza (para demandar su cumplimiento). Como resultado, dice Munn, existe una brecha entre los principios éticos y la práctica. Para él, en un mundo de suma cero, la obsesión con los principios de la IA no solo es inútil, sino también peligrosa, ya que desvía recursos humanos y financieros de enfoques más productivos. Un ejemplo de esto es cuando Google anunció la creación de un cuerpo independiente para revisar las prácticas éticas de IA de la compañía, pero este grupo no tendría poder real para vetar proyectos o detenerlos de manera significativa.

El llamado de Munn, en respuesta a estas limitaciones, es que se requiere un cambio de un enfoque basado en la adhesión universal a principios y reglas hacia un enfoque ético sensible al contexto. Esto implica un enfoque menos disciplinario y más emancipador, que permita a los actores morales actuar de manera autónoma y responsable, basándose en un conocimiento integral y en la empatía necesaria para tomar decisiones moralmente relevantes. Es decir, empoderar a quienes impulsan el desarrollo de la IA a tener estándares profesionales. En esto último concuerdo con Munn: el aspecto crucial de las críticas al desarrollo de la segunda ola de ética de la IA es que, aunque los principios han logrado cierta transversalidad en su adopción simbólica, no se ha logrado una apropiación de facto. Esta apropiación en la práctica es crucial, considerando que la legislación y regulación en materias de IA siempre estará incluso más retrasada que las consideraciones éticas.

De estas críticas consistentes en el tiempo, destaco que el paso a una tercera ola de la ética de inteligencia artificial debe exigir que los avances y desarrollos técnicos en el área se planteen desde la integración de la ética como parte del desarrollo de la IA, fortaleciendo estándares profesionales y estructuras reguladoras que enmarquen este proceso. Esto, sin embargo, implica necesariamente enseñar una ética aplicada a la inteligencia artificial para formar profesionales que usen los principios éticos, los cuales, aunque por sí solos pueden ser ineficientes, contextualizados y traducidos a prácticas concretas, pueden guiar mejores prácticas.

En respuesta a este escenario en el que aún nos encontramos, en la siguiente sección propongo como podemos navegar hacia una tercera ola de desarrollo de la ética de la IA desde la contextualización de la IA como sistema sociotécnico, comprendiendo esta transición desde el enmarcamiento (*Gestell*) heideggeriano.

2. Navegando hacia una tercera ola de la ética de IA: contextualización y el *desvelar* ético

Lo discutido anteriormente refleja la complejidad de avanzar en criterios éticos para el desarrollo e implementación de la IA. No basta con, simplemente, añadir la ética a la IA; una simple añadidura sería caer en lo que Johnson y Verdicchio llaman una falacia aditiva. Los autores han desafiado la noción simplista de "incrustar" (*embedding*) la ética en la IA, argumentando que, al añadir la ética a la tecnología; es decir, "Ética + IA = IA Ética," se comete una falacia al asumir erróneamente que los principios éticos pueden ser codificados directamente en la IA. Su argumento se basa en que, para lograr una adición de ese tipo, los campos de la IA y la ética deberían compartir características ontológicas, o sea, deben tener naturalezas compatibles, cosa que no tienen.

El origen ontológico de la IA tiene una base computacional, lo que implica que, para lograr dicha adición, la ética también debe ser "computacionalizada", capturando principios éticos en formulaciones matemáticas. Esta discrepancia pone énfasis en que cuando hablemos de ética en IA no debemos entenderlo como un proceso que implique incluir elementos o consideraciones éticas en la tecnología misma, sino que nos llama a ampliar nuestra manera de entender la relación que puede haber entre la ética y la tecnología. Desde esta limitación ontológica, una comprensión amplia de la IA como un sistema sociotécnico abre una dimensión de convergencia al considerar que las inteligencias artificiales no son solo un conjunto de herramientas computacionales, sino que operan dentro de redes complejas de relaciones humanas, normas sociales y prácticas organizativas, y es en esa dimensión donde puede haber una simbiosis.

Por esto, una de las tendencias que han surgido y a las cuales adscribo es definir la IA como un sistema sociotécnico, lo que implica concebirlo como una entidad que no opera de manera aislada, sino en contextos específicos; está inmersa en la sociedad, sujeta a sus normas y valores, y afectada por sus prácticas. Por lo tanto, cualquier consideración ética en la IA debe tener en cuenta el impacto y la interacción que la tecnología adopta y ejerce sobre el tejido social.

En la literatura, se han propuesto diversas definiciones y caracterizaciones de los sistemas sociotécnicos. En general, estos se comprenden como sistemas que dependen no solo del hardware técnico, sino también del comportamiento humano y de las instituciones sociales para su correcto funcionamiento (Davis *et al.*; Nickel). Según esta comprensión, Van de Poel (391) identifica que los sistemas sociotécnicos consisten en combinaciones de tres bloques de construcción básicos:

(i) Artefactos técnicos: objetos físicos que pueden cumplir y han sido diseñados para una cierta función técnica —tienen una naturaleza física¾.

(ii) Agentes humanos: personas que intencionan acciones e interactúan con artefactos técnicos.

(iii) Instituciones: reglas o normas sociales que deben ser seguidas por los agentes—prescriben expectativas de comportamiento moral¾.

A esta base fundacional, Van de Poel agrega dos características propias de los sistemas de IA: (iv) agentes artificiales y (v) normas técnicas. La idea subyacente aquí es que los agentes artificiales tienen propiedades que los distinguen de los artefactos técnicos tradicionales. Ahora bien, dado que las instituciones son construcciones sociales, nos dice Van de Poel, estas no pueden ser percibidas y seguidas directamente por agentes artificiales. Lo que sí pueden hacer es seguir un componente análogo que aplica a los sistemas de IA, a saber, las normas técnicas, que propician efectos causales-físicos en vez de en términos intencionales.

Basado en esto, Van de Poel articula que los sistemas de inteligencia artificial se definen como sistemas sociotécnicos que exhiben una complejidad aumentada en relación con los sistemas sociotécnicos convencionales, atribuible a su aptitud inherente para el aprendizaje y la adaptación. Esta propiedad distintiva tiene el potencial de inducir resultados imprevistos, los cuales plantean desafíos significativos en términos de previsión y regulación en comparación con sistemas tradicionales.

De ahí que la concepción de la inteligencia artificial como un sistema sociotécnico subraya la necesidad de examinar minuciosamente las interacciones que surgen entre los componentes tecnológicos (tales como agentes artificiales y normas técnicas) y los elementos sociales (incluyendo normativas, relaciones interpersonales y estructuras organizativas) en el proceso de diseño e implementación de sistemas basados en IA. Este análisis sociotécnico profundiza en la comprensión de los sistemas de IA más allá de su función computacional, adentrándose en su rol como configuradores de espacios sociales. Tal entrelazamiento resalta la importancia de considerar la tecnología de IA no solo como un ente de ejecución de tareas, sino como un agente activo en la reconfiguración de la interacción social y el conocimiento.

La consideración de la IA desde una perspectiva sociotécnica, en consecuencia, aboga por una comprensión holística de la tecnología, que incorpora a la IA en el tejido de las relaciones humanas y los marcos institucionales. Al reconocer que los sistemas de IA pueden tanto perpetuar como alterar las desigualdades

sociales, se pone de manifiesto la necesidad de desarrollar marcos de gobernanza tecnológica que sean éticamente robustos y eviten la falacia aditiva.

Estas consideraciones, no obstante, reflejan algo más que la mera necesidad de comprender la simbiosis y retroalimentación que hay entre técnica y sociedad para poder prevenir impactos negativos como un trato injusto o una violación a la privacidad. Una tecnología como la inteligencia artificial genera lo que Swierstra y Molder llaman "impactos suaves". Estos impactos suaves se refieren a las modificaciones sociales y personales que las tecnologías tienen en las personas, tales como la reevaluación de normas sociales y valores, la influencia en identidades sociales, incluso la reconfiguración de las expectativas sociales sobre bienestar. Por ejemplo, los sistemas de reconocimiento facial plantean nuevas expectativas sobre la privacidad y los límites de la vigilancia. También vemos cómo diagnósticos médicos asistidos por sistemas de IA pueden cambiar el valor que se le da a la tecnología y al rol humano en contextos de cuidado. Esto sugiere que una evaluación ética de la tecnología no debe limitarse a los impactos directos y evidentes, sino que también debe considerar estos efectos más sutiles y a largo plazo.

Para lograr un entendimiento más profundo de estos impactos suaves, propongo entender el paso a la tercera ola de la inteligencia artificial como una emancipación de los condicionamientos que la IA puede tener en nuestra sociedad, contextualizando su desarrollo e implementación para que nos sirva como una tecnología reveladora de las nuevas dinámicas éticas y epistémicas que implica la evolución de una sociedad que adopta y se tecnifica activamente con la IA. Esto implica usar los principios éticos que hemos construido como cimientos de una edificación que tenga ciertos mínimos éticos comunes, pero que se levante a través de un proceso hermenéutico que cuestione a la tecnología y nos permita tener un conocimiento y agencia sobre las verdades a las que accedemos gracias a ella.

2.1. Un proceso hermenéutico y un desvelar ético

Para embarcarnos en este proceso hermenéutico, primero debemos entender qué implica. Grunwald realiza una crítica a los métodos de evaluación tecnológica que se centran exclusivamente en las consecuencias de las tecnologías, argumentando que estos enfoques son insuficientes. Propone, en respuesta a su crítica, que también deberíamos examinar el conocimiento hermenéutico que tenemos disponible, es decir, la comprensión profunda e interpretativa sobre cómo las tecnologías son entendidas y significadas socialmente. Tomando esta definición de conocimiento hermenéutico, sugiero que parte de nuestras inquietudes epistemológicas respecto a una ética de la IA debe basarse en un proceso

iterativo para entender cómo significamos la tecnología, o sea, el significado social de la tecnología de IA. Así, al formar un espiral de entendimiento donde cada nuevo descubrimiento o interpretación modifica y enriquece nuestra comprensión inicial, podemos contextualizar los principios éticos que hemos generado. Ya no como simples directrices abstractas que han de "guiar" el actuar ético, sino como estructuras que cimientan nuestro proceso constante de cuestionar el rol de la tecnología en nuestra sociedad.

Esta aproximación hermenéutica, además, se complementa si la desenmarcamos en un sentido Heideggeriano. Para Heidegger, la esencia de la tecnología no es nada tecnológico, ya que la técnica no es un simple conjunto de herramientas; es más bien una forma de entender y relacionarse con el mundo (Heidegger). La técnica moderna, nos dice Heidegger, configura nuestra relación con el mundo de manera que todo se presenta como un recurso disponible para ser explotado, ayudándonos a desvelar (*Entbergen)* verdades. El desafío clave es no dejar la técnica en el dominio únicamente de lo técnico y, en cambio, cuestionarnos cómo podemos relacionarnos con la técnica de manera que nos permita acceder a una comprensión más rica de la verdad y del ser, en lugar de quedar atrapados en una visión meramente utilitaria del mundo.

Pero para alcanzar ese vínculo, debemos emanciparnos del enmarcamiento (*Gestell*) que la tecnología moderna tiene sobre nosotros, condicionándonos a un modo de comprensión del mundo y, por tanto, nuestra interacción con él. Según Heidegger, el enmarcamiento convierte todo en recursos disponibles para ser explotados. Si aplicamos esto al contexto de la IA, puede interpretarse como las tendencias a caer en ese tecnochovinismo o solucionismo tecnológico,[4] donde la tecnología de IA se ve como maximizadora y optimizadora de procesos y decisiones que impactan a los humanos, y no como una tecnología facilitadora para el bienestar humano. Esto también puede entorpecer nuestro acceso a diferentes conocimientos y verdades, ya que la IA, cuando no se contextualiza y no se entiende desde su dinámica inherentemente sociotécnica, puede caer fácilmente en vicios como la señalización virtuosa o la apariencia de autorregulación mencionados anteriormente. Un paso hacia la integración de la ética en el desarrollo de la tecnología de IA debe basarse en facilitar una relación auténtica de conocimiento, donde no nos condicionemos a la estructura tecnológica, sino que desvelemos sus dinámicas sociotécnicas para significarlas.

---

[4] El solucionismo tecnológico, como lo presenta Morozov tiende a simplificar problemas complejos, reduciéndolos a cuestiones que pueden ser resueltas mediante la tecnología, con aplicaciones, software o dispositivos. Esta simplificación, a menudo, pasa por alto factores socioculturales, históricos y políticos que son cruciales para entender y abordar dichos problemas de manera efectiva. Así, este solucionismo se caracteriza por un optimismo desmedido en la capacidad de la tecnología para mejorar la sociedad. Esto puede llevar a una dependencia excesiva de soluciones digitales y a la promoción de tecnologías como panaceas universales, tal y como ocurre con ciertas expectativas de integrar la IA indiscriminadamente.

Así, un proceso hermenéutico enfatiza cómo interpretamos y reinterpretamos tecnologías como la IA. A través de un "desvelar" continuo, empezamos a ver no solo lo que la IA puede hacer (su utilidad funcional) sino también lo que significa para la sociedad (su significado social y ético). Este "desvelar ético" lo defino así: entender cómo la IA puede cambiar nuestras relaciones sociales, nuestras estructuras de poder y nuestras concepciones de los diferentes principios que hemos establecido como su base normativa.

Esta visión, por tanto, invita a un cuestionamiento continuo sobre cómo la tecnología de IA puede ser diseñada y desplegada de manera que realmente beneficie y enriquezca la sociedad, no solo en términos de eficiencia y capacidad, sino también para fomentar una comprensión más profunda y crítica de nosotros mismos y del mundo que habitamos. Este paso a una tercera ola de la ética de inteligencia artificial debe, entonces, articular los principios existentes para que apoyen una reflexión más profunda sobre el desarrollo tecnológico, evitando integraciones superficiales que promuevan prácticas como la señalización virtuosa o la apariencia de autorregulación, mencionadas anteriormente. Esto implica comprometerse con un proceso continuo de reflexión y revisión sobre cómo la tecnología revela o encubre la verdad de nuestro ser en el mundo y cómo su rol en la sociedad influye en estos procesos epistémicos y éticos que determinan la coevolución de una sociedad enmarcada en esta tecnología.

Pero ¿qué implica, en la práctica, integrar un proceso de conocimiento hermenéutico que nos permita avanzar hacia una desvelación ética en el desarrollo de la IA? Primero, esto implica educar a los desarrolladores y usuarios de IA en un pensamiento crítico que contemple la tecnología como un actor que configura activamente nuestras realidades sociales y personales, demandando así una responsabilidad ética más profunda en su creación y uso. Uno de los ejes de desarrollo de esta nueva ola debe centrarse en la educación en la ética de la IA, que no debería limitarse a inculcar principios normativos prefijados, sino que debe fomentar un entendimiento dinámico de cómo las tecnologías revelan y ocultan aspectos del mundo que habitamos.

Además, es crucial integrar estos principios éticos en la práctica diaria del desarrollo de la IA. Esto no solo se logra a través de directrices y políticas, sino también mediante la implementación de mecanismos de revisión y ajuste constantes que permitan a las tecnologías adaptarse a nuevos entendimientos y desafíos éticos que surjan a lo largo de su ciclo de vida. Estos mecanismos deben ser transparentes y accesibles para los usuarios y desarrolladores, permitiendo un diálogo abierto sobre las decisiones éticas en el diseño y uso de la IA. Este diálogo, por cierto, debe surgir en un contexto cultural de reflexión continua. Esto incluye espacios de diálogo y crítica donde los implicados puedan discutir y revisar no solo los aspectos técnicos, sino también las implicancias sociales y

éticas de su trabajo. Estos foros deben ser inclusivos, incorporando una variedad de perspectivas, incluidas aquellas de las comunidades afectadas por estas tecnologías, para garantizar que la IA se desarrolle de manera justa.

Este paso hacia una tercera ola apunta a una contextualización de la IA como un sistema sociotécnico que influye y es influido por la sociedad y sus dinámicas. Es volver a responder esa pregunta por la técnica que nos planteaba Heidegger: al entender la esencia de la IA, podemos comprender cómo influye en nuestra forma de ser en sociedad y, por lo tanto, nos permite desvelar y emanciparnos de su influencia, transformándola en un medio para conocer y significar nuestra experiencia vital. Así, avanzar en este conocimiento hermenéutico de la IA nos permite desvelar la ética de la IA, al establecer una relación entre las expectativas normativas, las prácticas mediadoras y la tecnología misma, que enriquece nuestra condición humana y nuestro proceso de coevolución en una sociedad tecnificada por la IA.

Este enfoque hermenéutico no solo desafía la aplicación superficial de la tecnología, sino que nos exige repensar cómo los principios éticos están fundamentalmente entrelazados con el avance tecnológico. En la práctica, esto significa que las consideraciones éticas de la IA deben evolucionar de ser reactivas y situacionales a convertirse en proactivas e integrales en cada fase del desarrollo tecnológico. Tal enfoque exige un cambio de ver la ética de la IA meramente como un cumplimiento regulatorio o una estrategia de gestión de riesgos, para abrazarla como un pilar fundamental de la innovación y la filosofía de diseño. Esta transformación implica incorporar deliberaciones éticas desde el inicio de la conceptualización tecnológica, asegurando que cada paso del desarrollo esté alineado con la previsión ética y los valores sociales.

Además, la noción de desvelar éticamente dentro de la IA requiere un diálogo continuo entre múltiples partes interesadas, incluidos eticistas, tecnólogos, legisladores y el público. Este diálogo no debería limitarse solo a círculos académicos o de expertos, sino extenderse a la comunidad más amplia para democratizar el desarrollo de las tecnologías de IA. Al involucrar una variedad diversa de voces en la conversación, podemos anticipar y mitigar mejor los desafíos éticos que plantea la IA, asegurando que la tecnología avance de una manera no solo técnicamente competente sino también socialmente responsable y culturalmente sensible.

Por último, la aplicación de un enfoque hermenéutico en la ética de la IA también debería involucrar una evaluación y adaptación continuas de las directrices éticas a medida que la tecnología y sus implicaciones sociales evolucionen. Este proceso dinámico de calibración ética nos permite responder a nueva información, percepciones y desafíos, asegurando que la ética de la IA permanezca relevante y robusta. Al fomentar un marco ético adaptable, podemos facilitar una

integración más reflexiva de la IA en la sociedad, mejorando el potencial de estas tecnologías para servir como verdaderos instrumentos para el beneficio humano y social.

3. Comienzos del desvelar ético: desmitificar, enseñar e integrar

La combinación de un proceso de conocimiento hermenéutico y una revelación emancipadora del enmarcamiento de la tecnología nos invita a reconsiderar no solo cómo construimos tecnología, sino también para qué y para quién la construimos. Este marco no solo amplía nuestra comprensión de la tecnología como fenómeno social y ético, sino que también nos desafía a desarrollar tecnologías que verdaderamente enriquezcan la condición humana, reconociendo y respetando nuestra integridad y valores compartidos.

Así, al fomentar una ética de la IA sociotécnica que reconozca y respete la complejidad de la existencia humana, se invita a repensar el desarrollo tecnológico de la IA. Esta perspectiva sociotécnica requiere un enfoque de ética de IA que sea dinámico y adaptativo, capaz de responder a los cambios tecnológicos y sociales continuos. Tal enfoque no solo mejora la capacidad de la IA para servir al bien público, sino que también fomenta un desarrollo tecnológico más consciente y responsable en un sentido existencial, entendiendo que la tecnología tiene impactos duraderos y profundos en la sociedad.

En consecuencia, aunque las críticas a los principios éticos de IA sugieren su inutilidad por la falta de consecuencias tangibles, es crucial reconocer el papel significativo que han jugado en fomentar el debate y la conciencia ética en este ámbito. Los principios éticos en esta segunda ola de la ética de IA han funcionado como un catalizador para que la comunidad de la IA y la sociedad en general presten atención a las implicaciones morales y sociales del desarrollo de la IA. Si buscamos establecer un lenguaje común y objetivos éticos claros, estos principios son la base para que diferentes partes interesadas, desde desarrolladores hasta legisladores, tengan un punto de partida para discutir y evaluar las tecnologías emergentes. Por lo tanto, aunque puedan ser vistos como ineficaces en términos de ejecución directa, considero relevante enfatizar que su valor reside en elevar la ética de la IA a un tema de interés y discusión global, con directrices que nos permitan alcanzar ciertos mínimos comunes.

Además, las críticas a la ética de la IA, tanto por la falta de eficacia de los principios éticos como por las falacias aditivas cometidas, apuntan a la necesidad de contextualizar la IA como una tecnología sociotécnica, motivando el

desarrollo de metodologías interdisciplinarias que integren estos principios traduciéndolos a la práctica tecnológica. Operacionalizar intervenciones éticas, sin embargo, no puede alcanzarse solo desde formalizaciones computacionales (*e.g.*, métricas de *fairness*) o herramientas técnicas (*e.g.*, para auditorías de datos) ni solo desde principios o directrices (*e.g.*, disminuir la discriminación y aumentar la representatividad). Incorporar la ética activamente en el diseño y desarrollo de sistemas de IA, tal y como he sugerido aquí, implica una colaboración más estrecha entre eticistas, ingenieros, desarrolladores, científicos sociales y otros incumbentes (*stakeholders*) para crear un marco de trabajo que no solo sea teóricamente robusto, sino aplicable en contextos reales y que pueda desarrollarse desde un proceso hermenéutico fructífero. Hacer ética de inteligencia artificial que no caiga en la ineficiencia implica, entonces, el desarrollo de visiones más holísticas y colaborativas.

Por otro lado, es esencial fortalecer los currículos educativos en el campo de la IA para asegurar que los futuros profesionales comprendan que la ética no es un añadido superficial, sino una parte integral de la práctica profesional. Las percepciones de los desarrolladores respecto a la ética tienden a ser negativas o escépticas. Estudios exploratorios han evidenciado que, en algunos casos, se reconoce como una necesidad, pero que su experiencia con la ética es restrictiva, ya que impide avanzar en innovación, además no tiene directrices claras de cómo aplicarla a su trabajo (Arriagada *et al.*). En otros casos, los desarrolladores de IA entienden su trabajo desde una neutralidad; es decir, que ser un desarrollador no es ni ético ni no-ético, ya que sus decisiones son sobre eficacia, optimización y decisiones técnicas (Griffin *et al.*). En ambos estudios, se evidencia la adopción de una neutralidad peligrosa, que descontextualiza a la IA como una herramienta y no la entiende en su dimensión sociotécnica. En otras palabras, refleja posturas, perspectivas y prácticas en los desarrolladores de IA que los mantienen en un enmarcamiento que limita su entendimiento de esta tecnología y, por lo tanto, les impide identificar y significar su rol y experiencia en este camino de coevolución con la tecnología.

El llamado, por ende, al dar un paso a la tercera ola de la ética de IA es contextualizar y situar el conocimiento ético desde un proceso de conocimiento hermenéutico que tiene un rol emancipador y busca crear una cultura ética y de responsabilidad colectiva. Además, este enfoque hermenéutico implica entender la tecnología no como una entidad aislada, sino como parte de un tejido social y cultural más amplio, donde su desarrollo y aplicación están inevitablemente influidos por las estructuras de poder y las normas sociales existentes. Por lo tanto, es crucial que los desarrolladores de IA trabajen de cerca con comunidades, grupos de interés y expertos en ética para interpretar y entender las implicaciones éticas de la tecnología en diferentes contextos.

4. Conclusiones

En este artículo he presentado una serie de elementos que contribuyen a dilucidar el cambio paradigmático en el rol que tiene la ética de la IA en el desarrollo de esta tecnología para su tercera ola. La segunda ola de la ética de la IA, como se ha discutido, centra su enfoque en abordar preocupaciones prácticas emergentes vinculadas al uso cotidiano de tecnologías, a través de principios y definiciones como la transparencia, la privacidad y la justicia. Este enfoque, si bien es un avance respecto a las preocupaciones más teóricas de la primera ola, todavía puede estar impregnado de cierto solucionismo tecnológico, al intentar resolver problemas éticos complejos mediante soluciones tecnológicas "guiadas" por principios, como métricas de justicia para disminuir sesgos o sistemas de IA explicativos. Sin embargo, estas soluciones a menudo no consideran la totalidad de las estructuras sociales y políticas subyacentes, lo que puede llevar a la implementación de soluciones que no abordan las raíces de los problemas.

Concuerdo así, parcialmente, con las críticas a la segunda ola de la ética de IA, que apuntan a la insuficiencia de principios éticos abstractos y descontextualizados para abordar los retos prácticos y morales que presenta esta tecnología. Sin embargo, me parece que esta crítica no apunta a la brecha de fondo. Y es que los principios éticos no tienen como función última regular acciones y comportamientos. Para poder habilitar un contexto en el cual los principios éticos tengan una bajada aplicada, se requiere una comprensión profunda de las condiciones sociotécnicas en las que se despliegan los sistemas de IA y las implicancias éticas y epistémicas que esto tiene en contextos de desarrollo científico y social. De lo contrario, los principios éticos corren el riesgo de permanecer como ideales teóricos sin aplicabilidad real, incluso si hay regulaciones que exijan su uso.

Para esto, sostengo que la ética de la IA debe ser considerada como parte integral del diseño y la operación de sistemas tecnológicos, implicando una interacción constante con las normativas sociales y las dinámicas de poder. En el análisis sugiero que una aproximación hermenéutica, que contextualice y sitúe el conocimiento ético dentro de procesos dinámicos y cambiantes, es esencial para desarrollar prácticas que sean genuinamente éticas y efectivas. Esto resuena con la crítica hecha por Heidegger sobre el "*Gestell*" o enmarcamiento, donde la tecnología no solo se ve como un conjunto de herramientas, sino como un sistema que configura nuestra relación con el mundo, reduciendo todo a recursos para ser explotados. En este sentido, el solucionismo tecnológico puede ser visto como una manifestación del enmarcamiento, donde la tecnología se promueve como la solución universal, ignorando las dimensiones éticas, políticas y sociales más profundas.

Sugiero, por tanto, que la tercera ola de la ética de la IA puede ofrecer un camino de integración y un desvelar de la ética, al enfocarse en cómo la tecnología puede ser diseñada y utilizada de manera que respete y mejore la condición humana sin reducir a las personas y sus culturas a meros datos o recursos. Este enfoque más matizado y crítico permite una integración de la IA en la sociedad que es consciente de las complejidades humanas y éticas, y que busca soluciones que sean no solo efectivas desde el punto de vista tecnológico, sino también justas y sostenibles desde una perspectiva social y cultural.

Este cambio de paradigma, sin embargo, tiene también desafíos significativos en la estandarización y aplicación global de prácticas éticas en IA debido a variaciones culturales, legales y económicas que pueden influenciar esas develaciones contextuales. No obstante, esto puede ser también entendido como una oportunidad para desarrollar marcos éticos y metodologías éticas de trabajo que se ajusten a contextos locales mientras se mantienen alineadas con principios éticos transversales de la segunda ola.

A raíz del análisis propuesto en este trabajo, se llama a que investigaciones futuras se enfoquen en desarrollar métodos para la integración operativa de la ética en el ciclo de vida de los sistemas de IA, desde la concepción hasta la implementación y revisión desde las consideraciones sugeridas, para una integración más efectiva de los principios éticos en la práctica tecnológica. A través de un enfoque hermenéutico y la colaboración interdisciplinaria, podemos aspirar a desarrollar tecnologías de IA que sean éticas en un sentido robusto y no meramente por adición.

■ ■ ■ ■ ■

Referencias

Arriagada, Gabriela; López, Claudia y Mendoza, Marcelo. *Ethics of AI and IT.* CRC Press, Taylor and Francis, próxima publicación.

Beauchamp, Tom, y David DeGrazia. "Principles and Principlism". *ResearchGate*, 2004, pp. 55-74, https://doi.org/10.1007/1-4020-2127-5_3

Binns, Reuben. "Fairness in Machine Learning: Lessons from Political Philosophy." *Proceedings of Machine Learning Research*, vol. 81, 2018, pp. 1-11.

Boddington, Paula. *Towards a Code of Ethics for Artificial Intelligence*. 1st ed., Springer Cham, 2017. https://doi.org/10.1007/978-3-319-60648-4

Borenstein, Jason, *et al.* "AI Ethics: A Long History and a Recent Burst of Attention." *Computer*, vol. 54, n.° 01, Jan. 2021, pp. 96-102. https://doi.org/10.1109/MC.2020.3034950.

Bosk, Charles L. "Bioethics, Raw and Cooked: Extraordinary Conflict and Everyday Practice". *Journal of Health and Social Behavior*, vol. 51 Suppl, 2010, pp. S133-146. *PubMed*, https://doi.org/10.1177/0022146510383839.
Bostrom, Nick, y Eliezer Yudkowsky. "*The Ethics of Artificial Intelligence.*" 2014, pp. 316-34.
Broussard, Meredith. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, 2018.
Burrell, Jenna. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society*, vol. 3, n.° 1, June 2016, p. 2053951715622512. https://doi.org/10.1177/2053951715622512
Chouldechova, Alexandra. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". *Big Data*, vol. 5, n.° 2, 2017, pp. 153-63, https://doi.org/10.1089/big.2016.0047
Coeckelbergh, Mark. "Responsibility and the Moral Phenomenology of Using Self-Driving Cars". *Applied Artificial Intelligence*, vol. 30, n.° 8, Sept. 2016, pp. 748-57. https://doi.org/10.1080/08839514.2016.1229759
Danks, David, y Alex John London. "Algorithmic Bias in Autonomous Systems". *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, AAAI Press, 2017, pp. 4691-97, http://dl.acm.org/citation.cfm?id=3171837.3171944
Davis, Matthew C., *et al.* "Advancing Socio-Technical Systems Thinking: A Call for Bravery". *Applied Ergonomics*, vol. 45, n.° 2, Mar. 2014, pp. 171-80. *PubMed*, https://doi.org/10.1016/j.apergo.2013.02.009
Dencik, Lina, *et al.* "Towards Data Justice? The Ambiguity of Anti-Surveillance Resistance in Political Activism". *Big Data & Society*, vol. 3, n.° 2, Nov. 2016, p. 2053951716679678, https://doi.org/10.1177/2053951716679678.
Dennis, Louise, *et al.* "Formal Verification of Ethical Choices in Autonomous Systems". *Robotics and Autonomous Systems*, vol. 77, Mar. 2016, pp. 1-14, https://doi.org/10.1016/j.robot.2015.11.012.
Dignum, Virginia. "Ethics in Artificial Intelligence: Introduction to the Special Issue". *Ethics and Information Technology*, vol. 20, n.° 1, Mar. 2018, pp. 1-3. *Springer Link*, https://doi.org/10.1007/s10676-018-9450-z.
Etzioni, Amitai, and Oren Etzioni. "Incorporating Ethics into Artificial Intelligence". *The Journal of Ethics*, vol. 21, n.° 4, Dec. 2017, pp. 403-18. https://doi.org/10.1007/s10892-017-9252-2.
Floridi, Luciano, *et al.* "AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations". *Minds and Machines*, vol. 28, n.° 4, Dec. 2018, pp. 689-707. https://doi.org/10.1007/s11023-018-9482-5.
Griffin, Tricia A., *et al.* "The Ethical Agency of AI Developers". *AI and Ethics*, Jan. 2023, https://doi.org/10.1007/s43681-022-00256-3.
Grunwald, Armin. "The Objects of Technology Assessment. Hermeneutic Extension of Consequentialist Reasoning". *Journal of Responsible Innovation*, vol. 7, n.° 1, jan. 2020, pp. 96-112. https://doi.org/10.1080/23299460.2019.1647086.
Hagendorff, Thilo. "The Ethics of AI Ethics: An Evaluation of Guidelines". *Minds and Machines*, vol. 30, n.° 1, mar. 2020, pp. 99-120. https://doi.org/10.1007/s11023-020-09517-8.
Heidegger, Martin. *THE QUESTION CONCERNING TECHNOLOGY AND OTHER ESSAYS.* Translated by William Lovitt, Garland Publishing, Inc., 1977.
High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy AI*. Publications Office, 2019, https://doi.org/10.2759/346720.
Johnson, Deborah, y Mario Verdicchio. "*Ethical AI Is Not about AI - Communications of the ACM.*" 1 feb. 2023, https://cacm.acm.org/opinion/ethical-ai-is-not-about-ai/.
———. "*Ethics+AI Does not=Ethical AI:The Additivity Fallacy.*" University of Vienna.
Mittelstadt, Brent. "Principles Alone Cannot Guarantee Ethical AI". *Nature Machine Intelligence*, vol. 1, n.° 11, Nov. 2019, pp. 501-07. *arXiv.org*, https://doi.org/10.1038/s42256-019-0114-4.
Mittelstadt, Brent Daniel, *et al.* 'The Ethics of Algorithms: Mapping the Debate'. *Big Data & Society*, vol. 3, n.° 2, Dec. 2016, p. 2053951716679679, https://doi.org/10.1177/2053951716679679.
Morozov, Evgeny. *To Save Everything, Click Here. Technology, Solutionism, and the Urge to Fix Problems That Don't Exist*. Penguin Random House, 2014, https://www.penguin.co.uk/books/183571/to-save-everything-click-here-by-morozov-evgeny/9780241957707.
Munn, Luke. "The Uselessness of AI Ethics". *AI and Ethics*, vol. 3, n.° 3, Aug. 2023, pp. 869-77. https://doi.org/10.1007/s43681-022-00209-w.
Narayanan, Arvind. "*Tutorial: 21 Fairness Definitions and Their Politics*". https://www.youtube.com/watch?v=jIXIuYdnyyk. ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT).
Nickel, Philip J. "Trust in Technological Systems". *Norms in Technology: Philosophy of Engineering and Technology, vol. 9*, editado por M. J. de Vries *et al.*, Springer, 2013, https://philarchive.org/rec/NICTIT.
"Robotics: Ethics of Artificial Intelligence". *Nature*, vol. 521, n.° 7553, May 2015, pp. 415-18. https://doi.org/10.1038/521415a.

Swierstra, Tsjalling, y Hedwig te Molder. "Risk and Soft Impacts." *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, edited by Sabine Roeser *et al.*, Springer Netherlands, 2012, pp. 1049-66. https://doi.org/10.1007/978-94-007-1433-5_42.

van de Poel, Ibo. "Embedding Values in Artificial Intelligence (AI) Systems." *Minds and Machines*, vol. 30, n.° 3, Sept. 2020, pp. 385-409. https://doi.org/10.1007/s11023-020-09537-4.

van Wynsberghe, Aimee. "Sustainable AI: AI for Sustainability and the Sustainability of AI". *AI and Ethics*, vol. 1, n.° 3, Aug. 2021, pp. 213-18. https://doi.org/10.1007/s43681-021-00043-6.

■ ■ ■ ■ ■