https://doi.org/10.17163/abyaups.141.3

Implicaciones éticas de la inteligencia artificial

Alex Darío Estrada García
Universidad Nacional de Educación (UNAE)
alex.estrada@unae.edu.ec
https://orcid.org/0000-0001-5278-8221

Jandry Geomar Chuni Gaona
Universidad Nacional de Educación (UNAE)
jandry.chuni@unae.edu.ec
https://orcid.org/0009-0006-5618-076X

Introducción

La eclosión de la inteligencia artificial generativa (IA) insta a discutir sobre las prácticas históricamente asentadas en la cotidianeidad del ser humano, una de ellas son las cuestiones relacionadas con la ética. En este escenario, es trascendente establecer reflexiones críticas sobre las implicaciones éticas que podrían presentarse en el contexto social a partir de la utilización de la IA. Terminologías como responsabilidad, decisiones, transparencia, seguridad, privacidad, vulnerabilidad, sesgos cognitivos demandan un abordaje filosófico urgente para establecer diálogos que guíen a la dilucidación de explicaciones plausibles relacionadas con el diseño, uso y evaluación de programas de IA en la cotidianeidad del ser humano.

Dignum (2018, p. 2) propone tres niveles de integración de la ética en la IA: *ethics by desing* (integración técnica de la ética en el algoritmo),

ethics in desing (revision y análisis de las implicaciones éticas durante el diseño de programas), *ethics for desing* (códigos de conducta y estándares éticos para programadores). Hablar de ética en el contexto de la IA demanda pensar el proceso de diseño de los programas, al igual que escuchar las opiniones e intereses de las partes en cuestión. Es preciso tener presente que en varias narrativas con respecto a las implicaciones de la IA en la vida de los humanos se posicionan desenlaces catastróficos en un futuro próximo; en muchas ocasiones, esto corresponde a valoraciones morales descontextualizadas de los avances tecnocientíficos (Zamora, 2021).

El abordaje de las cuestiones expuestas tendrá lugar desde el relativismo moral metaético. Esta posición filosófica expone que los juicios morales a menudo dan directivas prácticamente contradictorias y ninguno de los juicios puede demostrar ser racionalmente superior al otro. La verdad moral no es absoluta, es relativa, por consecuente, no se deberá entender los juicios morales de modo literal (Tasioulas, 2010; Zamora, 2021; Gowans, 2021).

Algunas interrogantes que conducen el desarrollo del presente capítulo son ¿qué es la IA? ¿qué creemos que es la IA? ¿en qué puede convertirse, de acuerdo con las proyecciones humanas, la IA? ¿las máquinas necesitan tomar decisiones éticas? ¿la ética en el diseño de la inteligencia artificial es fundamental? ¿cómo definir cánones éticos para el desarrollo, implementación y evaluación de programas de IA?

Se destaca que la IA trae muchos beneficios a los diferentes campos del conocimiento (Holmes *et al.*, 2021), así como a la cotidianidad de los seres humanos. Actualmente, el panorama es distinto al de la Revolución Industrial, cuando las máquinas eran un complemento del humano, mientras que ahora lo sustituyen (Brynjolfsson y McAfee, 2014). Esta sustitución da lugar a un sinnúmero de especulaciones relacionadas con el quehacer humano y de las máquinas en un futuro cercano. Un cambio de época trae consigo nuevas narrativas sobre posibilidades soñadas que se cultivan en el lenguaje común y se posicionan como verdades. Por

ejemplo, la inminente transformación tecnológica actual eran los relatos de la ciencia ficción de los siglos pasados.

Las nuevas realidades, sobre todo las virtuales, traen "consecuencias no deseadas de la tecnología" (Coeckelbergh, 2021, p. 85). Es decir, escapan de la predicción de los sistemas técnicos y su impacto en los diversos escenarios de aplicación. Por lo expuesto, más allá de una IA distópica, general, superinteligente, es decir, un escenario en que se alcanzaría una singularidad tecnológica; la preocupación debería ser centrarse en pensar y construir la ética en la IA, la cual según Coeckelbergh (2021) debería "[…] preocuparse del cambio tecnológico y su impacto en las vidas de los individuos, pero también de las transformaciones en la sociedad y en la economía" (p. 20). En consonancia con lo expuesto en este apartado, se propone como objetivo analizar y describir las implicaciones éticas de la IA y las repercusiones en los contextos sociales.

Inteligencia artificial generativa

Las definiciones de inteligencia artificial (IA) varían considerablemente entre los autores, lo que refleja las diversas perspectivas desde las cuales se aborda el concepto. Chesterman (2020) critica la ambigüedad del término "IA"; señala que su significado se diluye en un espacio de interpretaciones y aplicaciones, lo que puede dificultar su regulación y comprensión pública. Esta ambivalencia se manifiesta en la forma en que distintos sectores, desde la academia hasta la industria y la política, abordan la IA, cada uno con sus propias prioridades y preocupaciones.

Desde un enfoque pragmático, para los propósitos legales y regulatorios es eficaz definir la IA en función de casos de uso y riesgos específicos. De esta manera, los legisladores podrían enfocarse en delimitar la tecnología en contextos particulares donde se presenten riesgos o impactos significativos (Vilaça *et al.,* 2024). Esto como respuesta a la no existencia de una definición universalmente aceptada. Se sugiere que, en lugar de intentar abarcar toda la extensión del concepto de IA, que

es inherentemente vasto y en evolución, es más útil considerar cómo se aplica en situaciones concretas que involucran las decisiones autónomas, los datos o la seguridad.

De acuerdo con lo expuesto, el desafío radica no solo en encontrar un término común que pueda ser aceptado por todas las partes interesadas, sino también en abordar las implicaciones éticas y sociales que emergen en cada contexto, asegurando que las definiciones utilizadas no solo sean técnicamente precisas, sino también responsables y sensibles a las preocupaciones del público. Esto requerirá un esfuerzo colaborativo entre investigadores, reguladores y la sociedad en general para garantizar que la evolución de la IA beneficie a todos de manera equitativa.

Wang (2019) examina las diferentes maneras en que se ha definido la IA a lo largo del tiempo y las compara. Argumenta que se puede entender la IA desde cinco perspectivas: estructura, comportamiento, capacidad, función y principios. Esta clasificación resalta que las definiciones pueden variar según si intentan replicar el funcionamiento cerebral (estructura), su comportamiento observable (comportamiento), su habilidad para resolver problemas (capacidad) o las funciones específicas que desempeña (función). Además, esta diversidad en las definiciones no solo refleja el avance técnico en el campo, sino también la evolución de la comprensión humana sobre la cognición y la inteligencia, lo que se traduce en debates filosóficos sobre la naturaleza de la inteligencia artificial y su contraste con su contraparte humana.

En un enfoque más amplio, Sheikh *et al.* (2023) conceptualizan la IA como una "tecnología de sistema" que influye en todos los niveles de la sociedad, comparándola con revoluciones tecnológicas pasadas como la electricidad o el motor de combustión. A diferencia de Chesterman (2020) y Wang (2019), estos autores destacan no solo el aspecto técnico de la IA, sino también las profundas implicaciones sociales, económicas y éticas que conlleva. Su argumento sugiere que la integración de los sistemas informáticos autónomos generativos contemporáneos no es un mero fenómeno técnico.

Por ejemplo, Bruderer (2020) explica que la Agencia Europea de Defensa la define como la capacidad de los algoritmos para seleccionar la mejor opción entre un amplio conjunto de posibilidades, mientras que un Centro de Excelencia del Gobierno de EE. UU. la describe como sistemas capaces de realizar tareas de manera autónoma en contextos variables. El desarrollo histórico resalta el viaje de la IA desde un concepto teórico hasta convertirse en una herramienta tangible que permea el área industrial en el diario vivir. A pesar de enfrentar períodos de estancamiento, la IA ha resurgido con avances significativos que han revitalizado el interés en su investigación y aplicación. Programas como Deep Blue, que venció a campeones de ajedrez que respondió preguntas sobre una variedad de temas, simbolizan hitos en este resurgimiento. Las aproximaciones al desarrollo de la IA incluyen herramientas simbólicas, sistemas expertos y razonamiento basado en casos (CBR), que permiten a los sistemas aprender de experiencias previas y adaptarse a nuevas situaciones (Bruderer, 2020). Este enfoque multifacético posibilita que la IA no solo realice tareas específicas, sino que también mejore su rendimiento a medida que interactúa con su entorno, un principio fundamental en su evolución.

Las herramientas no simbólicas comprenden sistemas multiagente, formados por unidades de software que se adaptan y colaboran para abordar problemas complejos. Ha sido fundamental para las aplicaciones actuales de la IA, utilizando arquitecturas y algoritmos que aprenden de los datos a través de redes neuronales artificiales y técnicas de aprendizaje automático, que han evolucionado para permitir a las máquinas aprender y generalizar de manera más eficiente (Ali *et al.*, 2023). Esta evolución ha facilitado el surgimiento de aplicaciones prácticas multidisciplinar hasta la automoción, donde la capacidad de aprender de grandes volúmenes de datos es crucial.

Este avance ha sido impulsado por el acceso a grandes volúmenes de datos y mejoras en las capacidades computacionales, que han permitido entrenar modelos más complejos y precisos. La esencia de la IA generativa radica en algoritmos que aprenden de ejemplos no supervisados, con redes

neuronales como su estructura fundamental. Estas redes, organizadas en capas, ajustan los pesos de las neuronas para minimizar las funciones de costo, demostrando su capacidad de generalizar y aprender a partir de ejemplos (Ali *et al.*, 2023). Esta capacidad de aprendizaje no solo se traduce en una mayor precisión, sino también en una versatilidad que permite a la IA generativa adaptarse a diferentes tipos de contenido y aplicaciones.

La IA generativa comprende métodos y aplicaciones que crean contenido similar al producido por humanos, como texto, imágenes, música o código. Estas tecnologías se entrenan con grandes volúmenes de ejemplos reales, usualmente de manera no supervisada, para aprender características del contenido a generar. Con instrucciones del usuario, generan nuevos contenidos coherentes con su entrenamiento. Programas como *Gemini, ChatGPT* y *Copilot* han revolucionado la generación automática de contenido, permitiendo respuestas detalladas ante consultas precisas. Esto ha transformado la interacción con la información y ha abierto nuevas vías para la creatividad e innovación (Casar-Corredera, 2023).

Los enfoques bajo observación son importantes para la inteligencia artificial de tipo generativa. En el modelo supervisado se aprende bajo datos nominados, lo que facilita tareas como clasificación y regresión, aprendiendo a mapear entradas a salidas específicas. En cambio, el aprendizaje no supervisado utiliza datos no etiquetados, permitiendo que los modelos identifiquen patrones sin guía explícita. Esto es esencial para la generación de nuevas muestras en IA generativa, como en *GANs* (*Generative Adversarial Networks*) y *autoencoders* variacionales (Kingma y Welling, 2013). La capacidad de aprender sin supervisión es crucial en contextos donde los datos etiquetados son costosos o difíciles de obtener.

Las aplicaciones de IA generativa trascienden la generación de texto. Herramientas como *DALL-E* y *MidJourney* crean imágenes 'originales' en diversos estilos artísticos, mientras que otras se enfocan en la composición musical y la generación de efectos a largo plazo. La definición de estas herramientas varía según el contexto, reflejando las amplias capacidades de las IA generativas en distintos ámbitos creativos (Gupta *et al.*, 2024).

Esta flexibilidad y adaptabilidad subrayan el potencial no solo como herramientas de producción, sino también como motores de innovación en el arte y el diseño.

En el diseño de productos, por ejemplo, puede generar prototipos innovadores basados en tendencias de mercado y preferencias del consumidor, lo que optimiza el proceso de desarrollo y reduce el tiempo de comercialización (Suphavarophas *et al.,* 2024). En la programación de videojuegos, permite la creación de niveles y escenarios únicos que se adaptan a las decisiones de los jugadores, generando experiencias personalizadas y dinámicas (Hu *et al.,* 2023). Además, en campos como la biotecnología y la medicina, la IA generativa puede simular estructuras moleculares y proponer nuevos fármacos, acelerando el descubrimiento y desarrollo de tratamientos efectivos (Vamathevan *et al.,* 2019). Estos ejemplos ilustran cómo la IA generativa no solo mejora la eficiencia, sino que también impulsa la creatividad y la innovación en disciplinas altamente especializadas.

La adaptación de la arquitectura de los modelos de IA generativa a las características específicas de cada dominio es esencial para maximizar su efectividad. Este es el caso de los diseños de productos, los modelos pueden incorporar algoritmos de optimización para evaluar la viabilidad funcional y estética de los diseños generados (Bongiorno *et al.*, 2022). En el caso de los videojuegos, se pueden utilizar arquitecturas como redes neuronales convolucionales (CNN) para procesar y generar elementos visuales, mientras que redes recurrentes (RNN) pueden ayudar en la creación de narrativas dinámicas (Hu *et al.*, 2023). Esta capacidad de personalizar la arquitectura del modelo según las necesidades del dominio permite que la IA generativa sea una herramienta versátil y poderosa en la innovación y creación de contenido en múltiples campos. Esta personalización también implica un desafío adicional: la idea de un amplio espectro de comprensión características del dominio para que los modelos puedan ser diseñados de manera efectiva considerando diversos lineamientos éticos.

La IA generativa permite crear contenido personalizado que se adapta a las necesidades de cada usuario, utilizando algoritmos que analizan datos sobre comportamientos y elecciones previas. Esta personalización mejora la experiencia del usuario, ya que ofrece recomendaciones y soluciones más relevantes, desde recomendaciones de productos hasta contenidos en plataformas de *streaming* (Ricci *et al.*, 2010). Sin embargo, la personalización propone retos significativos en base de la privacidad y ética, puesto que implica la administración y almacenamiento de muchos datos. Esto puede dar lugar a preocupaciones sobre cómo se utilizan los datos, quién tiene acceso a ellos y cómo se protegen los derechos del usuario (Hagendorff, 2020).

La necesidad de un enfoque ético y transparente en la implementación de la IA generativa es crucial para garantizar que los beneficios de la personalización no se vean eclipsados por violaciones de privacidad y desconfianza en la tecnología (Möllmann *et al.*, 2021). Este aspecto ético es fundamental, dado que una implementación irresponsable puede socavar la confianza del usuario y limitar la adopción de tecnologías basadas en IA generativa.

Sesgo algorítmico, responsabilidad y transparencia

En el mundo contemporáneo, la IA está posicionando una visión de intento de modelado de la sociedad desde formas profundas y omnipresentes. Esta tecnología tiene el poder de transformar campos educativos, económicos y médicos; sin embargo, con grandes capacidades vienen grandes responsabilidades. En particular, existen preocupaciones crecientes sobre el sesgo algorítmico y sobre cómo se garantiza el desarrollo e implementación de los mecanismos de inteligencia artificial. En este epígrafe de explora estos temas, subrayando la importancia de un enfoque ético y responsable en el uso de la IA.

Introducción a las causas y consecuencias del sesgo algorítmico

El sesgo algorítmico es una problemática crítica que surge cuando los modelos de IA reflejan o amplifican prejuicios preexistentes en los datos o en las estructuras de los algoritmos. Estos sesgos no son una simple falla técnica, sino una manifestación de cómo se desarrolla y entrena la tecnología en un mundo que ya está lleno de desigualdades y prejuicios (Aparicio-Gómez y Cortéz, 2024).

El sesgo algorítmico tiene múltiples causas. Un principal desencadenante es la presencia de datos sesgados, los algoritmos de IA aprenden a partir de datos preexistentes y, si estos datos contienen prejuicios, el resultado será un modelo igualmente sesgado. Por ejemplo, esto sucede en el entrenamiento de los sistemas de IA que generan texto; si el modelo de IA fue entrenado con datos que contienen estereotipos relacionados con género, etnia, cultura, religión, es muy probable que el sistema de IA genere textos que refuerzan estos estereotipos, de esta forma se podría ocasionar algún tipo de discriminación, exclusión o desconfianza en la adopción de los programas de IA. Otro ejemplo claro se da en los sistemas de contratación automatizada; si el historial de contratación de una empresa muestra una preferencia por contratar a hombres en lugar de mujeres, un algoritmo entrenado con esos datos tenderá a replicar ese comportamiento, descartando a candidatas calificadas por prejuicio inherente en los datos (Kheiri *et al.*, 2024).

Otra causa importante es el diseño del algoritmo. Los desarrolladores están sujetos a introducir sesgos involuntariamente al elegir cómo se deben ponderar ciertas características o al definir la estructura del modelo. Los algoritmos son productos humanos y, como tales, pueden reflejar las limitaciones y perspectivas del equipo de desarrollo. En los sistemas de reconocimiento facial, por ejemplo, se ha demostrado que los algoritmos tienen mayores tasas de error al identificar personas de color en comparación con personas de tez blanca. Esta situación se debe, en

parte, a que los datos utilizados en el entrenamiento tienden a tener una mayor representación de personas blancas, mientras que la diversidad étnica es insuficiente (Etxeberria, 2024). En el ámbito de la publicidad, los algoritmos de segmentación pueden discriminar de manera indirecta al mostrar ciertos anuncios solo a un segmento específico de la población, limitando las oportunidades para otros. Esto afecta de manera exponencial el acceso equitativo a la información, servicios y productos.

Sobre *cómo mitigar el sesgo algorítmico en la IA*, inicialmente, se sostienen que requiere un enfoque integral que aborde las causas desde diferentes perspectivas (Kaufman, 2024). En primer lugar, es esencial contar con conjuntos de datos que sean representativos y diversos. Un buen punto de partida es realizar una revisión y curación cuidadosa de los datos, para garantizar que no incluyan patrones sesgados de comportamiento o estereotipos sociales (Raghavan *et al.*, 2020). Por ejemplo, en el estudio de la *ciencia de datos* se evidencian esfuerzos por desarrollar técnicas de muestreo para equilibrar el peso de ciertos grupos dentro de los conjuntos de datos de entrenamiento.

Otra estrategia es la auditoría de los algoritmos enfocada en la evaluación constante de los modelos para identificar y mitigar posibles sesgos. Esta práctica puede llevarse a cabo mediante auditorías internas o, preferiblemente, por auditores externos que puedan ofrecer una perspectiva imparcial. Es importante tomar en cuenta que los desarrolladores deben contar con herramientas que posibiliten analizar la equidad de los resultados generados por sus modelos, como indicadores que evalúan la disparidad entre los diferentes grupos demográficos.

A partir de los aportes de Burrell (2016) menciona que la transparencia en los algoritmos es fundamental, ya que no solo permite a los usuarios entender el proceso de toma de decisiones, sino que también les ofrece la oportunidad de identificar y corregir sesgos que podrían contribuir a la perpetuación de desigualdades en las decisiones automatizadas. Finalmente, es esencial promover la diversidad en los equipos de desarrollo. Los desarrolladores y científicos de datos de diversas culturas,

géneros y perspectivas pueden identificar y abordar problemas que por ciertas cuestiones están sujetas a pasar por alto.

La responsabilidad en la IA implica que las empresas y los desarrolladores asuman la obligación ética y moral de los impactos que sus sistemas tienen en la sociedad. La responsabilidad debe ser un principio rector en todas las fases de la vida de un sistema de IA: desde el diseño, pasando por el entrenamiento, la implementación, el uso y la evaluación. Los principios de responsabilidad requieren que las empresas tengan la responsabilidad de las decisiones y acciones de sus algoritmos. Esto significa que no se pueden escudar en la autonomía tecnológica como una forma de evadir la rendición de cuentas.

Una problemática común de este tipo es atribuir la responsabilidad de una decisión errónea a la máquina, ignorando que la programación y el uso del sistema son responsabilidad humana (Mittelstadt, 2019). Para evitar esto, las empresas están en la obligatoriedad de establecer mecanismos claros para identificar quién es responsable en cada etapa del desarrollo y uso del sistema. Esto se puede lograr mediante la asignación de roles específicos de revisión ética dentro del equipo de desarrollo, y documentando cada decisión técnica que impacte la equidad y justicia del sistema. La transparencia es un componente vital para lograr la responsabilidad en el trabajo y fomento de la creación de la IA. La transparencia algorítmica implica que los procesos que llevan a la adquisición de la voluntad por parte de la inteligencia artificial los cuales deben ser comprensibles para todos los usuarios. Esto incluye la posibilidad de acceder a información sobre qué datos se utilizaron para entrenar el sistema y cómo estos influyen en sus resultados.

Un ejemplo donde la falta de transparencia ha generado controversia es el caso de los algoritmos de crédito, que determinan la viabilidad crediticia de las personas (Coeckelbergh, 2020a; 2022). Si los usuarios no comprenden porqué se les negó un crédito, se fomenta la desconfianza en el sistema financiero. En el libro *La sociedad del desconocimiento* de Daniel Innerarity (2022), se explica que el problema radica en la 'caja negra' de

los algoritmos, donde ni siquiera los desarrolladores entienden con precisión cómo un sistema de aprendizaje profundo llega a una conclusión (Pasquale, 2015). La transparencia debe incluir no solo la posibilidad de revisar los resultados, sino también hacer públicos los datos y métodos utilizados, de manera que los expertos externos puedan analizarlos y realizar sugerencias.

Conforme con lo expuesto, se plantean algunas ideas que invitan a pensar en estrategias direccionadas a la mitigación del sesgo algorítmico. En la literatura científica se debate la necesidad de implementar estrategias efectivas desde la fase de diseño la IA. Esto incluye la recopilación de datos diversificados y representativos que reflejen la realidad de la población en su conjunto. López (2023) argumenta que la inclusión de diferentes datos de entrenamiento es muy importante para evitar el sesgo en los algoritmos que se usan. Esto puede incluir técnicas de corrección de sesgos que ajusten los resultados algorítmicos para garantizar que no se favorezca a un grupo sobre otro (Coeckelbergh, 2022). Los estudios en esta área están en constante progreso y desarrollan métodos innovadores para abordar el sesgo en tiempo real. Además, es importante fomentar la colaboración interdisciplinaria en la evolución de la IA. Involucrar a expertos en ética, sociología, derecho y otras disciplinas en el proceso de diseño puede ayudar a identificar y abordar preocupaciones relacionadas con el sesgo y la justicia social.

Los desafíos que presenta la IA demanda entender la profunda interrelación entre sesgo algorítmico, responsabilidad y transparencia. La presencia de sesgos en los algoritmos afecta directamente la calidad y equidad de las decisiones, lo cual se traduce en injusticias y desigualdades para aquellos afectados por los resultados del sistema de IA. Por tanto, la responsabilidad implica que los desarrolladores y empresas deben ser conscientes de los sesgos existentes, hacerlos visibles mediante la transparencia y trabajar activamente para reducirlos (Gallifant *et al.*, 2024).

Para enfrentar los problemas derivados del sesgo algorítmico, la transparencia es clave. Sin información clara sobre los datos y métodos

utilizados, se torna casi imposible que los reguladores, académicos y el público en general puedan verificar si un sistema es justo y equitativo. Al promover la transparencia, las empresas tecnológicas están fomentando la confianza en la tecnología y sus nuevos productos. Además, la responsabilidad también implica trabajar activamente para la mejora continua de los sistemas, incorporando mecanismos de auditoría y control que aseguren la equidad, y cuidando que los sistemas de IA no se utilicen para perpetuar o amplificar desigualdades existentes.

Regulación y marcos éticos para la IA

Para este cometido es fundamental el análisis y regulación con el propósito de establecer un marco que garantice el desarrollo ético y responsable de la IA. Dado que los gestores de IA tienen el potencial de impactar a nivel mundial, es importante que las políticas públicas ofrezcan un marco de referencia que defina límites y establezca responsabilidades claras. Algunos países han avanzado en la creación de regulaciones para la IA. La Unión Europea, por ejemplo, ha propuesto la "Ley de Inteligencia Artificial", que se basa en un enfoque de regulación proporcional al riesgo que representa un sistema (Binns, 2022). Los sistemas de IA que tienen el potencial de afectar derechos fundamentales o causar daños significativos están sujetos a requisitos más estrictos de transparencia y evaluación de riesgos.

Las recomendaciones propuestas por la UNESCO (2021), también ofrecen directrices para el desarrollo responsable de la IA. Estos marcos se centran en garantizar que la IA sea beneficiosa para la sociedad, respetando los derechos humanos y promoviendo el bienestar. Sin embargo, la implementación de estos principios depende de la voluntad política y empresarial de cumplir con los compromisos adquiridos. Es necesario que la ética no sea solo una declaración de intenciones, sino que esté integrada en el proceso de diseño e implementación (Floridi y Cowls, 2019).

Es inminente la dificultad de aplicar un marco ético global que responda a todas las interrogantes que pueden surgir en los diversos

escenarios de aplicación de la IA; esto debido, en primera instancia, a la diversidad de definiciones asignadas al término *ética*. Históricamente se ha evidenciado los acuerdos y desacuerdos en alcanzar definiciones plausibles; sin embargo, lo que se tiene claro es que el uso de este término es relativo a un conjunto de personas que comparten valores, creencias, dogmas, ideologías. Por tal razón, alcanzar un marco ético en el que se refleje el consenso de qué valores y prioridades deben ser atendidas en el desarrollo de la IA, es poco alcanzable. Otra razón que imposibilita alcanzar un marco ético general que exponga ciertos principios es la complejidad de los sistemas de IA, en algunos ámbitos son impredecibles el alcance de estos sistemas. A esto se suma el avance vertiginoso de la tecnología, los cambios continuos harían obsoleto cualquier intento de generalización de lineamientos o principios éticos generales.

A pesar de las diversas dificultades en el contexto ético, puede implementarse medidas útiles para el desarrollo adecuado de la IA. Por ejemplo, a nivel estatal, como ya se evidenció en el inicio de este apartado, es oportuno crear regulaciones y leyes para normar el desarrollo y uso de la IA. Así mismo, es de suma importancia la cooperación internacional destinada a intercambiar avances tecnológicos y marcos éticos que posibiliten mitigar determinados riesgos. En cuanto a las instituciones que desarrollan sistemas de IA están llamados a transparentar el funcionamiento y proyectar los posibles impactos que pueda llegar a tener; para ello, es fundamental contar con códigos de ética propios de la institución.

De acuerdo con lo expuesto, se plantea un interrogante que invita a pensar el desarrollo, funcionamiento y evaluación de los sistemas de IA: ¿cuál es el rol de la filosofía en el desarrollo tecnológico? Diversas son las razones por las que se debe subrayar la filosofía de la tecnología, sobre todo para constituir un marco ético. Por ejemplo, posibilita un análisis conceptual de términos como 'ética', 'inteligencia', 'responsabilidad', 'conciencia' que se emplean sin concisión ni precisión, lo cual dificulta el debate sobre el desarrollo tecnológico y sus dimensiones. De igual forma, la filosofía proporciona herramientas para el análisis de dilemas

éticos que eclosionan en el ámbito en mención. Estas son dos de varias funciones que podría tener la filosofía de la tecnología, sobre todo para el fortalecimiento de un marco conceptual y metodológico encaminado a dilucidar cuestiones relacionadas con la ética de/en IA.

Privacidad de los datos

Una de las características propias de la tercera década del siglo XXI es la cantidad de información que generan los seres humanos a partir de su interacción con el medio social y natural. Las sofisticadas Tecnologías de la Información y Comunicación (TIC) son medios para crear, difundir, acceder a una vasta información dando paso a una transferencia de datos globalizada, comúnmente se lo ha llamado *big data*. Por lo tanto, la sociedad está inmersa en un mundo *datificado*, preguntarse por la privacidad de los datos es crucial. En el contexto de la IA, los datos son la materia prima para el entrenamiento de los modelos de lenguaje masivos (large language models [LLM]) que están constituidos por redes neuronales artificiales (Wang *et al.*, 2023). En tal sentido, la IA se favorece de los LLM por la capacidad de estos para el procesamiento de datos y, a través de la conjugación de algoritmos, ofrecen respuestas en lenguaje natural a preguntas planteadas por los humanos (Estrada-García y Narváez, 2024).

Es oportuno precisar en el tipo de información del que están constituidos los gigantescos bulos de datos. Como se había mencionado en la parte introductoria del presente capítulo, actualmente, la tecnología es el campo central al que el ser humano se ha mudado, cada vez más son las actividades digitales que alimentan los bulos con información de todo tipo. Por ejemplo, Coeckelbergh (2021) argumenta que "la IA, y en particular las aplicaciones de aprendizaje automático que operan con el *big data* a menudo implican la recogida y el uso de información personal" (p. 85). El entrenamiento de los programas de IA se da con todo tipo de información, sea esta contrastada o no; por ello, la predicción de resultados es cuestionable tanto por el tipo de información utilizada como por la configuración de algoritmos, puesto que pueden crearse a partir de sesgos.

La IA se respalda de campos del conocimiento como la ciencia de datos, específicamente, la estadística se constituye como uno de los aliados más importantes en el entrenamiento de algoritmos. El procesamiento de datos se da a partir de lo conocido como *aprendizaje automático* que, dependiendo de los objetivos, puede ser supervisado o no. Como consecuencia se obtiene la predicción de resultados a partir de la identificación de patrones en una gran cantidad de datos, este proceso en el lenguaje común se lo conoce como minería de datos. El procedimiento mencionado se complementa con el *aprendizaje por refuerzo*, es cual está direccionado para retroalimentar la predicción que el programa de IA ofreció.

Un caso importante que hay que destacar es el relacionado con las redes sociales —cualquiera que sea— con la privacidad de los datos personales que insertan los usuarios. Evidentemente, en algunas ocasiones estas aplicaciones solicitan consentimiento a los usuarios para utilizar los datos personales con el propósito de 'ofrecerle una mejor experiencia'. Sin embargo, ¿los usuarios conocen lo que estas redes sociales hacen con sus datos? No es posible acceder transparentemente a información sobre cómo se están utilizando los datos que producen los seres humanos diariamente. Esta práctica de las redes sociales se constituye como un mecanismo en que se utiliza al ser humano como mano de obra gratuita, es más se podría decir que se da una explotación, todo esto genera datos que serán analizados y procesados por la IA.

> Un uso ético de la IA requiere que los datos sean recogidos, procesados y compartidos de una forma que respete la privacidad de los individuos y su derecho a saber lo que ocurre con ellos, al acceso de los mismos, a objetar su recogida o procesamiento y a saber que se están recogiendo y procesando y (si procede) que están expuestos a la decisión tomada por una IA. (Coeckelbergh, 2021, p. 86)

Los datos generados a partir de las acciones humanas en entornos mediados por la tecnología, se utilizan para entrenar a los programas de IA. Por ello, es trascendente discutir los principios éticos bajos los cuales puede darse la utilización de los datos, configuración de algoritmos,

retroalimentación. La preocupación humana por una ética para/de la IA nace de lo que creemos que es y en lo que pueda convertirse la IA. Por ejemplo, se ha mencionado que la IA tiene el potencial de convertirse en una inteligencia superior a la que posee el ser humano. Al respecto, la científica Boden (2016) explica que, si bien está de acuerdo en la eminente configuración de una IA general, superinteligente, sostiene que esto no ocurrirá en la práctica; por lo tanto, la IA, comparada con la inteligencia humana, es menos prometedora de lo que mucha gente piensa y aún no conseguirá superarla, al menos no sucederá en las próximas décadas.

Manipulación, vigilancia y totalitarismo

La obtención, procesamiento y análisis de grandes cantidades de información insta a discutir sobre la manipulación, vigilancia y totalitarismo en diversos escenarios. La información personal de los usuarios ha sido utilizada por programas de IA para definir preferencias relacionadas con productos y servicios ofertados por empresas. Las redes sociales gracias a sus características propias acaparan datos personales de los usuarios que son usados, en principio, para una supuesta personalización de la información que estas redes sociales ofrecen al público. Sin embargo, está la posibilidad de que los datos personales sean usados para fines políticos, económicos o de otra índole.

Indudablemente, tecnologías como la internet y los *smartphones* han cambiado totalmente la realidad histórica del ser humano y no siempre estas tecnologías aportan solo cuestiones positivas, sino también negativas (Tollon, 2024). De hecho, las nuevas tecnologías instauran formas de manipulación, vigilancia y totalitarismo diferentes a las tradicionales, formas que quizá sean poco evidentes para las masas, no obstante, efectivas para determinados fines. Los *smartphones,* acompañantes diarios de los seres humanos, se constituyen a partir de aplicaciones que generan datos continuamente, a través de estas se da una suerte de vigilancia indirecta.

En los escenarios de la política, la IA en equipo con las redes sociales puede convertirse en una herramienta utilizada para la manipulación de

las sociedades. Por ejemplo, mediante la utilización de los datos generados por las redes sociales se puede influir en campañas políticas. Por las cuestiones expuestas y otras más, se habla del uso ético de las tecnologías ante la preocupación por las decisiones personales o de grupos de poder que no están prescritas por la ley (Etzioni y Etzioni, 2017).

De acuerdo con la visión de Coeckelbergh (2011) "[u]na forma de analizar y evaluar lo que las tecnologías de la información hacen y podrían hacer a las personas y a la sociedad, es utilizar el enfoque de capacidad como un marco normativo-ético" (p. 81). El enfoque referido hace hincapié en la comprensión de una ética de la tecnología centrada en las capacidades humanas. Un enfoque funcional que propone revisar el proceso de cómo la tecnología aporta a lo que las personas son capaces de hacer de acuerdo con un contexto social, cultural y tecnológico particular.

De conformidad con lo expuesto, la preocupación por la ética, específicamente, en el escenario de la IA recae en dos cuestiones. La primera, consiste en lo que la IA puede hacer, y segundo, en lo que el ser humano está en la capacidad de hacer con la IA. Pensar la primera cuestión lleva a la aceptación de una IA autónoma, si bien esta idea sea, al menos por unos años, utópica. Mientras que la segunda cuestión es la que demanda de la atención prioritaria con el propósito de plantear la integración de la ética en la configuración técnica de los programas de la IA, una revisión y análisis de las implicaciones éticas durante el diseño y códigos de conducta y estándares éticos para programadores (Morley *et al*., 2021; 2023).

Shanahan (2015) explica que la IA puede llevar a los humanos a ser "menos capaces de pensar o de decidir por sí mismos lo que hacer" (p. 170). Esta cuestión desencadena un sinnúmero de preocupaciones como la infantilización de los usuarios ante la posibilidad de que la IA asuma tareas propias del ser humano, como las cognitivas. De ser este el caso, los usuarios consumen combinaciones de algoritmos y la actividad cognitiva se verá menos forzada. Lo presentado son escenarios en los que la manipulación está presente, de igual forma, la explotación de datos personales (Coeckelbergh, 2019; 2020b).

Una cuestión que se debe discutir es si ¿las máquinas equipadas con inteligencia artificial necesitan ser capaces de tomar decisiones morales?, de ser este el caso ¿existe responsabilidad moral en las decisiones que toman las máquinas? Al respecto, Etzioni y Etzioni (2017) explica que las máquinas con IA no necesitan tomar decisiones éticas, siempre y cuando, estén en entornos gobernados por la ley, puesto que estas regulan y guían el comportamiento adecuado, al igual que establecerían los límites de acción de las máquinas y de los humanos que configuran a estas.

Por último, en un mundo interconectado es evidente la susceptibilidad de los programas y dispositivos tecnológicos a sufrir hackeos o cualquier tipo de ataque malintencionado. De igual forma está la vigilancia y el control de los gobiernos a las sociedades a través de la IA. Como respuesta, es fundamental promover una ética positiva para la privacidad, la transparencia y la rendición de cuentas. Todos son, una u otra forma, vulnerables a la tecnología, desde los problemas ligados a la seguridad hasta la creciente dependencia humana a las maquinas con IA; es decir, la creciente designación de tareas a las máquinas.

Conclusión

La vida de los seres humanos cambia constantemente por la introducción de programas, dispositivos tecnológicos cada vez más sofisticados, lo que demanda de una reflexión crítica a fin de discutir y exponer implicaciones positivas o negativas que podrían originarse de la introducción de programas de IA en la cotidianeidad humana. La IA ha dotado a las máquinas funciones que hacen posible la interacción mediante lenguaje natural con el ser humano, posibilidad que hace que muchas tareas históricamente complejas, se hagan más sencillas.

Los humanos nos hemos convertido en seres tecnológicos, a diario se utiliza un sin número de programas, dispositivos tecnológicos para la consecución de objetivos de la vida diaria, profesional, entre otras cuestiones dependiendo de los intereses. Cabe destacar que no todos los avances tecnológicos son positivos, también se dan aspectos que pueden

ser catalogados como negativos. Por ejemplo, la privacidad de los datos que los usuarios generen y son utilizados para el entrenamiento de programas de IA a partir del aprendizaje automático. De la misma manera, la transparencia en la configuración de algoritmos y el aprendizaje por refuerzo que retroalimenta al programa de IA para mejorar las respuestas.

Es importante mencionar que el sesgo algorítmico, la responsabilidad y la transparencia son elementos fundamentales en el desarrollo de sistemas de IA que respeten los principios éticos y promuevan el bienestar social. El sesgo algorítmico puede tener graves repercusiones en distintos ámbitos —desde la justicia hasta el acceso a servicios financieros— y su mitigación requiere un enfoque sistémico que abarque desde la calidad de los datos hasta la estructura de los algoritmos y la diversidad en los equipos de desarrollo. Dependerá de la capacidad del ser humano para enfrentar estos desafíos éticos con honestidad y dedicación. Si se logran implementar de manera adecuada prácticas de transparencia y responsabilidad y se trabaja activamente para mitigar los sesgos, se podrá crear sistemas de IA justos, inclusivos y capaces de mejorar la vida de las personas, sin importar su origen o condición.

La responsabilidad y la transparencia son esenciales para asegurar que los desarrolladores y empresas asuman las consecuencias de sus creaciones. Estos principios no solo promueven la confianza del público en los sistemas de IA, sino que también permiten la rendición de cuentas y la implementación de mejoras para garantizar la equidad. El desafío de reducir los sesgos, implementar la transparencia y garantizar la responsabilidad recae en los desarrolladores, las empresas y, en última instancia, los reguladores y la sociedad en general.

Una conjugación entre la IA, acceso a internet y las redes sociales ofertará más oportunidades para la manipulación, aparte que diversos mecanismos complejizan este problema. Por ejemplo, investigadores de la Universidad de Washington (BBC, 2017) crearon con IA un discurso de Barack Obama. Acciones como estas pueden desencadenarse con mayor frecuencia, la creación y difusión de noticias falsas elaboradas por IA,

utilizando la identidad de personajes influyentes en la sociedad puede ocasionar impactos catastróficos en diversos campos.

La mezcla entre hechos y ficción cada vez es más frecuente, el problema de la verdad y posverdad entra en cuestión; saber distinguir será uno de los desafíos en de la era de la IA. La manipulación, vigilancia y totalitarismo disfrazados de libertad acecha a una sociedad tecnológica, gobernada por máquina y programas que cada vez escapan del "control" humano.

Referencias bibliográficas

Ali, R., Hussain, A., Nazir, S., Khan, S. y Khan, H. U. (2023). Intelligent Decision Support Systems - An Analysis of Machine Learning and Multicriteria Decision-Making Methods. *Applied Sciences*, *13*(22), 12426. https://doi.org/10.3390/app132212426

Aparicio-Gómez, O. y Cortés Gallego, M. (2024). Desafíos éticos de la inteligencia artificial en la personalización del aprendizaje. *Revista Interamericana de Investigación Educación y Pedagogía 17*(2), 377-392. https://doi.org/10.15332/25005421.10000

BBC (17 de julio de 2017). El falso Barack Obama creado con inteligencia artificial capaz de hablar como si fuera el original. https://bit.ly/46OPLWO

Binns, R. (2022). Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance*, *16*(1), 197-211. https://doi.org/10.1111/rego.12358

Boden, M. (2016). *AI: Its Nature and Future.* Oxford University Press.

Bongiorno, A., De La Torre, L. y Caffaro, F. (2022). Generative design in the digital era: A review. *Journal of Computing and Information Science in Engineering, 22*(1), 1-20.

Bruderer, H. (2020). *Milestones in analog and digital computing* (3rd ed.). Springer.

Brynjolfsson, E. y McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & company.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2-12. https://doi.org/10.1177/2053951715622512

Casar-Corredera, J. (2023). Inteligencia artificial generativa. *Anales de la Real Academia de Doctores*, *8*(3), 475-489.

Chesterman, S. (2020). Artificial intelligence and the limits of legal personality. *International & Comparative Law Quarterly*, *69*(4), 819-844. https://doi.org/10.1017/S0020589320000366

Coeckelbergh, M. (2011). Human development or human enhancement? A methodological reflection on capabilities and the evaluation of information technologies. *Ethics and Information Technology*, *13*, 81-92. https://doi.org/10.1007/s10676-010-9231-9

Coeckelbergh, M. (2019). Artificial intelligence: some ethical issues and regulatory challenges. *Technology and regulation*, *2019*, 31-34. https://doi.org/10.26116/techreg.2019.003

Coeckelbergh, M. (2022). *Robot ethics*. MIT Press.

Coeckelbergh, M. (2020a). *AI ethics*. MIT Press.

Coeckelbergh, M. (2020b). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, *26*(4), 2051-2068. https://doi.org/10.1007/s11948-019-00146-8

Estrada-García, A. y Narváez, J. P. (2024). ChatGPT y la superficialidad del conocimiento: implicaciones académicas y éticas en el siglo XXI. *Yachana Revista Científica*, *13*(2), 19-36. https://doi.org/10.62325/10.62325/yachana.v13.n2.2024.911

Etxeberria Guridi, J. F. (2024). The use of artificial intelligence (AI) systems for remote biometric identification in publicly accessible spaces in the European AI law. *Actualidad Jurídica Iberoamericana*, *21*, 528-565.

Etzioni, A. y Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, *21*, 403-418. https://doi.org/10.1007/s10892-017-9252-2

Floridi, L. y Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1), 1-18. https://doi.org/10.1162/99608f92.8cd550d1

Gallifant, J., Bitterman, D., Celi, L., Gichoya, J., Matos, J., McCoy, L. y Pierce, R. (2024). Ethical debates amidst flawed healthcare artificial intelligence metrics. *npj Digital Medicine, 7*(243). https://doi.org/10.1038/s41746-024-01242-1

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. y Courville, A. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Gowans, C. (2021). *Moral Relativism.* The Stanford Encyclopedia of Philosophy. Edited by Edward N. Zalta. https://bit.ly/4gYRdKY

Gupta, P., Ding, B., Guan, C. y Ding, D. (2024). Generative AI: A systematic review using topic modelling techniques. *Data and Information Management, 8*(2), 100066. https://doi.org/10.1016/j.dim.2024.100066

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and machines*, *30*(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B. y Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, *2*, 504-526. https://doi.org/10.1007/s40593-021-00239-1

Hu, Z., Ding, Y., Wu, R., Li, L., Zhang, R., Hu, Y. y Fan, C. (2023). Deep learning applications in games: a survey from a data perspective. *Applied Intelligence*, *53*(24), 31129-31164. https://doi.org/10.1007/s10489-023-05094-2

Innerarity, D. (2022). *La sociedad del desconocimiento*. Galaxia Gutenberg.

Kaufman, D. (2024). Logic and foundations of artificial intelligence and society's reactions to maximize benefits and mitigate harm. *Unisinos Journal of Philosophy, 25*(1), 1-13. https://doi.org/10.4013/fsu.2024.251.10

Kheiri, M. A., Qashta, N., & Aljaradat, D. I. (2024). Authenticity of using artificial intelligence systems in proving electronic evidence. *Pacific Journal of Legal & Social Sciences, 23*(1), 37-52. https://doi.org/10.57239/PJLSS-2025-23.1.004

Kingma, D. P. y Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

López, D. M. (2023). Retos de la inteligencia artificial y sus posibles soluciones desde la perspectiva de un editorialista humano. *Biomédica, 43*(3), 309-314.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501-507. https://doi.org/10.1038/s42256-019-0114-4

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J. y Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of AI ethics. *Minds and Machines*, *31*(2), 239-256. https://doi.org/10.1007/s11023-021-09563-w

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M. y Floridi, L. (2023). Operationalising AI ethics: barriers, enablers and next steps. *AI & Society*, 38, 411-423. https://doi.org/10.1007/s00146-021-01308-8

Möllmann, N. R., Mirbabaie, M. y Stieglitz, S. (2021). Is it alright to use artificial intelligence in digital health? A systematic literature review on ethical considerations. *Health Informatics Journal*, *27*(4), 14604582211052391. https://doi.org/10.1177/14604582211052391

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Harvard University Press.

Raghavan, M., Barocas, S., Kleinberg, J. y Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of

the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469-481).

Ricci, F., Rokach, L. y Shapira, B. (2010). Introduction to recommender systems handbook. In Ricci, F., Rokach, L., Shapira, B., Kantor, P. (eds). *Recommender systems handbook* (pp. 1-35). Springer. https://doi.org/10.1007/978-0-387-85820-3_1

Shanahan, M. (2015). *The technological singularity.* The MIT Press.

Sheikh, H., Prins, C., Schrijvers, E. (2023). AI as a System Technology. In: Mission AI. Research for Policy (pp. 84-134). Springer. https://doi.org/10.1007/978-3-031-21448-6_4

Suphavarophas, P., Wongmahasiri, R., Keonil, N. y Bunyarittikit, S. (2024). A Systematic Review of Applications of Generative Design Methods for Energy Efficiency in Buildings. *Buildings*, *14*(5), 1311.

Tasioulas, J. (2010). Relativism, realism, and reflection. *Inquiry, 41*(4), 377-410. https://doi.org/10.1080/002017498321706

Tollon, F. (2024). Technology and the situationist challenge to virtue ethics. *Science and Engineering Ethics*, *30*(10), 1-17. https://doi.org/10.1007/s11948-024-00474-4

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura – UNESCO (2021). *Recomendación sobre la ética de la inteligencia artificial.* Conferencia General de la UNESCO, Paris del 9 al 24 de noviembre de 2021. https://bit.ly/3KsXvpW

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G. y Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug discovery*, *18*(6), 463-477. https://doi.org/10.1038/s41573-019-0024-5

Vilaça, M., Lopes, I. y Ferro, M. (2024). AI beyond a new academic hype: an interdisciplinary theoretical analytical experiment (computational, linguistic and ethical) of an AI tool. *Unisinos Journal of Philosophy, 25*(1), 1-14. https://doi.org/10.4013/fsu.2024.251.12

Wang, F.-Y., Miao, Q., Li, X., Wang, X. y Lin, Y. (2023). What does ChatGPT say: The DAO from algorithmic intelligence to linguistic intelligence. *IEEE/CAA Journal of Automatica Sinica, 10*(3), 575-579. https://doi.org/10.1109/JAS.2023.123486

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1-37. https://doi.org/10.2478/jagi-2019-0002

Zamora Bonilla, J. (2021). *Contra apocalípticos: Ecologismo, Animalismo, Posthumanismo.* Shackleton Books.