JÓVENES + IA: PERSPECTIVAS DE UNA GENERACIÓN
JÓVENES + IA: PERSPECTIVAS DE UNA GENERACIÓN

JÓVENES + IA
PERSPECTIVAS DE UNA GENERACIÓN

Ensayos sobre el presente y los desafíos de la inteligencia artificial para los derechos humanos

Tecnológico de Monterrey
Escuela de Humanidades y Educación

IL-DA

Jóvenes + IA: perspectiva de una generación

Publicado en diciembre de 2025 por la Iniciativa Latinoamericana por los Datos Abiertos (ILDA) en colaboración con la Escuela de Humanidades y Educación del Tecnológico de Monterrey, Campus Guadalajara.

Índice

Agradecimientos.....4
Prólogos.....**5**

IA y derechos humanos: ¿qué desafíos tenemos en América Latina?
Sesgos y discriminación en la inteligencia artificial, la censura digital en América Latina como dilema ético emergente.....13
Injusticia estructural y las decisiones automatizadas en México y América Latina.....22
La sociedad civil como pilar de la gobernanza de la IA en América Latina: transparencia, rendición de cuentas y desempeño democrático.....33
En torno a las decisiones autónomas de la Inteligencia Artificial: libertad, juicio y responsabilidad.....39

Los dilemas éticos de la IA
El Riesgo de la Neutralidad Algorítmica: Ética y Poder en las Decisiones Automatizadas.....48
El Espejo Digital: La Igualdad de Género en la IA y Nuestro Futuro Social.....59
Entre la caja negra y la dignidad humana: desafíos éticos de la IA en la justicia.....69
IA sin reglas: ¿avance o amenaza?.....81
Deberes y Dilemas: La Gobernanza Ética de la IA y los Derechos Humanos.....90
Hacia un Futuro Responsable: Ética, Datos e Inteligencia Artificial en Debate.....98
Reflejos de poder: políticas de una sociedad fragmentada.....119
Derechos Humanos e Inteligencia Artificial: Análisis de cómo los sistemas de IA pueden respetar, proteger o violar los derechos humanos.....130
Humanidad reducida a datos.....139
Dilemas éticos en el uso de chatbots emocionales como acompañamiento humano.....160

Agradecimientos

Quiero expresar mi más profundo agradecimiento a la Escuela de Humanidades y Educación del Tecnológico de Monterrey, Campus Guadalajara, y en particular a su director regional, Fernando A. Mora Dávila, por su invaluable apoyo, entusiasmo y visión para hacer posible esta iniciativa. Su compromiso con la formación crítica y humanista de las y los estudiantes fue fundamental para que este proyecto pudiera materializarse.

Extiendo también un reconocimiento especial al equipo de personas revisoras, cuyo rigor, dedicación y mirada experta enriquecieron de manera significativa la calidad de los trabajos seleccionados. Su labor fue esencial para garantizar un proceso académico sólido, transparente y cuidadoso.

Agradezco igualmente al Índice Global de Inteligencia Artificial Responsable (GIRAI) por las facilidades otorgadas para el uso de sus resultados como base de análisis en esta convocatoria. Su apertura y colaboración permitieron que las y los estudiantes contaran con evidencia robusta y actualizada para nutrir sus reflexiones.

Finalmente, reconozco de manera especial el trabajo de Aremi González y Violeta Belver, del equipo de ILDA, cuyo acompañamiento constante, coordinación y dedicación hicieron posible el desarrollo de este proceso de principio a fin. Su compromiso con la promoción de debates éticos, inclusivos y situados en América Latina fue un pilar fundamental de este proceso.

A todas y todos, mi más sincero agradecimiento por contribuir a que este proyecto se convirtiera en una realidad.

Gloria J. Guerrero Martinez,
Directora Ejecutiva, ILDA

Prólogos

Gloria J. Guerrero Martinez, Directora Ejecutiva de ILDA

Desde la **Iniciativa Latinoamericana por los Datos Abiertos (ILDA)** de la mano de la **Escuela de Humanidades del Tecnológico de Monterrey, Campus Guadalajara**, nos honra presentar este proyecto, resultado de la convocatoria universitaria para escribir sobre ética, datos e inteligencia artificial. Esta obra reúne las reflexiones de estudiantes que, desde distintas disciplinas y contextos, asumieron el desafío de analizar los impactos éticos de la inteligencia artificial en nuestras sociedades.

Cuando lanzamos esta convocatoria, nuestro objetivo era claro: abrir un espacio para que las y los jóvenes universitarios exploraran los dilemas y oportunidades que surgen en un mundo cada vez más impulsado por sistemas automatizados. Nos parece fundamental promover análisis éticos situados y comprometidos con la realidad latinoamericana y es por ello que se propuso utilizar los resultados de la primera edición del *Índice Global de Inteligencia Artificial Responsable (GIRAI)* como detonador de las preguntas y reflexiones.

Lanzado en 2024, el **Índice Global de IA Responsable** es la primera herramienta diseñada para evaluar el estado de la *IA responsable*. La primera edición del Índice evaluó la situación de la IA responsable en 138 países, con más de la mitad de ellos ubicados en el Sur Global. En términos generales, el Índice encontró que los países están rezagados de forma significativa en la provisión de marcos seguros y éticos para el despliegue de la IA en todos los sectores. Al examinar dimensiones sociales, técnicas y de políticas clave, esta herramienta permitió identificar brechas, tendencias y oportunidades para avanzar en el desarrollo de sistemas de IA alineados a los derechos humanos.

El propósito de este ejercicio fue que las y los estudiantes conocieran estos resultados y reflexionaran sobre las implicaciones que esta realidad tiene para su futuro y las sociedades latinoamericanas. Una mejor comprensión sobre cómo está avanzando el Sur Global en la implementación y uso de la IA, delinea los próximos pasos hacia la puesta en acción de políticas y prácticas concretas. Los artículos que conforman

este libro cumplen ese propósito. Cada autor y autora desarrolló un análisis profundo en torno a temas como la discriminación algorítmica, la igualdad de género, la transparencia, la explicabilidad de la IA, y las implicaciones éticas de delegar decisiones a sistemas automatizados.

A este ejercicio sumamos las reflexiones de expertos y expertas que desde distintas organizaciones en América Latina trabajan esta agenda. La combinación de miradas es un ejercicio valioso y necesario.

Creo firmemente que involucrar a la juventud en estas reflexiones es también una oportunidad invaluable para fortalecer la comprensión crítica sobre el impacto de la tecnología en nuestra vida cotidiana. En un contexto donde los datos son el insumo fundamental para el diseño y funcionamiento de sistemas y algoritmos, es imprescindible reconocer que el uso y clasificación de estos datos no son neutros: están atravesados por decisiones políticas, estructuras económicas, desigualdades sociales y relaciones de poder históricas. Entender esta no neutralidad permite cuestionar qué voces se incluyen y cuáles quedan fuera, qué realidades se priorizan y qué efectos generan estas ausencias en los modelos que influyen en políticas públicas, mercados y dinámicas sociales. Promover esta conciencia en las y los jóvenes no solo enriquece su capacidad de análisis, sino que contribuye a formar una ciudadanía más crítica y comprometida con la construcción de tecnologías que reflejen los valores democráticos, la equidad y la justicia social que nuestra región demanda.

El **Comité Evaluador**, compuesto por personas expertas como Nicolás Grossman, Mariel García-Montes, Mariana Rozo-Paz, y los profesores Rodrigo Esparza, Fernando A. Mora Dávila y Natalia Rocha, revisó cuidadosamente cada propuesta. Agradezco profundamente su compromiso e interés en este proyecto.

La selección final refleja el rigor académico, la originalidad de los enfoques y la pertinencia de los análisis para los debates actuales sobre gobernanza digital, datos abiertos y ética de la IA en América Latina y el Caribe. La calidad de los artículos destaca no solo por su solidez teórica, sino también por su capacidad para proponer soluciones, cuestionar narrativas y abrir nuevas rutas de reflexión.

Este libro es, en ese sentido, un testimonio del compromiso y la visión de una nueva generación de estudiantes que entiende que el futuro de la inteligencia artificial debe construirse desde principios éticos claros, desde la defensa de los derechos humanos y desde una perspectiva regional que reconozca

nuestras desigualdades, necesidades y aspiraciones.

En ILDA creemos firmemente que estos aportes son indispensables para avanzar hacia tecnologías más justas, transparentes y responsables. Agradezco también a todas las personas que participaron en la convocatoria, a quienes integraron el comité evaluador, al equipo gestor de este proyecto y especialmente a quienes hoy forman parte de esta publicación.

Esperamos que este libro inspire nuevas discusiones, investigaciones y acciones que fortalezcan la gobernanza responsable de los datos y la inteligencia artificial en América Latina. Que estas páginas sean una invitación a seguir construyendo, de manera colectiva, un futuro digital ético, diverso y verdaderamente inclusivo. Es por ello que también se decidió publicarlo en un formato de acceso libre y digital, bajo el espíritu de que esta discusión no quede solo en las aulas o en los espacios de expertos, sino que sea un detonante de reflexiones abiertas, colaborativas e inclusivas.

Bienvenidas y bienvenidos a esta obra que celebra el pensamiento crítico y el compromiso ético de la tecnología en nuestra región.

Fernando A. Mora Dávila, Director, Región Centro Occidente, Campus Guadalajara, Tec de Monterrey

El año 2025 ha estado marcado por una mezcla de expectación, entusiasmo y especulación en torno a la inteligencia artificial. Quizá nunca antes la humanidad había invertido tantos recursos en una sola tecnología bajo el ya conocido argumento de que todo se hace "en nombre del progreso". En el discurso público se repite que la Inteligencia Artificial traerá nuevos trabajos, aumentará la productividad y resolverá aquello que como sociedad no hemos sabido resolver. Ha sido un año de *tecnoptimismo*: una confianza casi ciega en que esta nueva herramienta solucionará todos nuestros problemas. Y, en medio de ese entusiasmo, se han dejado de lado, o se han tratado de forma superficial, debates urgentes que deberían formar parte central de la conversación pública.

Daron Acemoglu y Simon Johnson, en *Poder y Progreso*, advierten que vivimos en tiempos cada vez más elitistas, dominados precisamente por ese optimismo tecnológico que oscurece las desigualdades existentes. El progreso, recuerdan, no es un proceso automático: las innovaciones actuales vuelven a concentrar beneficios en un grupo pequeño de emprendedores e inversionistas, mientras que la mayoría obtiene muy poco y carece de capacidad real de decisión. Si queremos una relación distinta, más justa e inclusiva con la tecnología, necesitamos transformar también la base del poder social. Eso solo es posible si surgen voces divergentes, argumentos críticos y organizaciones capaces de contrarrestar las narrativas dominantes.

Por eso, abrir espacios para que hablen las y los jóvenes se vuelve tan relevante. Si para nosotros los adultos la Inteligencia Artificial es una tecnología disruptiva, para ellos será un dispositivo que moldeará su futuro. De ahí la importancia de amplificar sus perspectivas, de ofrecer un espacio donde puedan pensar, cuestionar y expresar cómo imaginan el mundo que les tocará vivir.

En el Tecnológico de Monterrey pensamos a las humanidades digitales como esenciales para cumplir este rol fundamental de espacios para el pensamiento crítico. Las humanidades digitales nos ayudan a situar la Inteligencia Artificial dentro de los marcos

culturales, históricos, éticos y sociales que la hacen comprensible. Frente a una visión técnica que presenta la IA como algo inevitable o neutral, las humanidades digitales permiten ver cómo los algoritmos moldean identidades, influyen en decisiones y reconfiguran relaciones de poder. Al combinar métodos computacionales con preguntas sobre derechos humanos, justicia, agencia, representación, sesgos, memoria y autonomía, este campo nos permite entender no solo cómo funciona la IA, sino qué implicaciones tiene para la vida y su cohabitar. Desde ahí podemos promover debates más informados, impulsar diseños tecnológicos responsables y abrir espacios donde diversas comunidades puedan imaginar futuros más justos y humanos.

Este libro es resultado directo de esas discusiones sobre ética, Inteligencia Artificial y datos. En un mundo obsesionado con chips, semiconductores, GPUs y centros de datos más grandes que treinta campos de fútbol, suele olvidarse algo básico: sin datos no existe la inteligencia artificial. Y en los datos, en cómo se obtienen, cómo se procesan y cómo se interpretan, se juegan algunos de los dilemas éticos más profundos de nuestro tiempo.

A partir de la reflexión ético-filosófica y utilizando el *Global Index on Responsible AI*, las y los estudiantes exploraron sesgos en el uso de datos para entrenar modelos algorítmicos, riesgos de la automatización de decisiones, tensiones entre gobernanza y derechos humanos en la IA, dinámicas de poder, el papel de la sociedad civil, los peligros de asumir la neutralidad algorítmica y la reducción de la experiencia humana a datos.

Incluso desde una mirada filosófica más amplia, como la que propone Heidegger, estas transformaciones adquieren un matiz inquietante. Un mundo organizado por la Inteligencia Artificial puede acercarnos a lo que él describe como el Ge-Stell: una forma moderna de entender la realidad que convierte todo, incluso al ser humano, en un recurso disponible y calculable. Cuando nuestra experiencia cotidiana se traduce en datos y patrones, existe el riesgo de empezar a vernos a nosotros mismos bajo esa misma lógica de eficiencia y predicción. En este sentido la IA no es solo una herramienta que procesa información; también influye en cómo entendemos el mundo, en cómo nos entendemos y relacionamos dentro de él.

En este libro se reúnen las reflexiones de expertas en el tema y de doce estudiantes de los campus Guadalajara, Puebla y Cuernavaca del Tecnológico de Monterrey. Acompañados por sus profesores, desarrollaron análisis rigurosos desde diversas metodologías de la ética, dando forma a textos que combinan mirada crítica, sensibilidad humanística y comprensión técnica.

Agradezco especialmente al profesor Rodrigo Esparza Parga (Campus Puebla) y al profesor Omar Cerrillo Garnica (Campus Cuernavaca) por su compromiso y por guiar a sus estudiantes en este proceso de reflexión profunda.

La selección de los textos incluidos en esta obra se dio en varias etapas: primero, a partir de un filtro inicial realizado por los profesores de la materia de Ética e Inteligencia Artificial; y posteriormente mediante la revisión de un comité integrado por profesoras y profesores del Tecnológico de Monterrey, así como por representantes de ILDA y actores de la sociedad civil. Este proceso colegiado permitió asegurar la calidad académica de los trabajos y, al mismo tiempo, reconocer la diversidad de perspectivas que aportaron las y los estudiantes.

Finalmente, esta obra fue posible gracias a la colaboración con ILDA, cuyo apoyo, en particular a través del *Global Index on Responsible AI*, permitió detonar muchas de las reflexiones presentes en estas páginas y, sobre todo, fortalecer la convicción de que las voces jóvenes deben ocupar un lugar central en la conversación sobre el futuro tecnológico.

Que este libro sea un atisbo de esperanza en un mundo cada vez más incierto, y que el pensamiento crítico y filosófico siga alumbrando caminos para construir respuestas más humanas ante los retos tecnológicos que enfrentamos, avanzando hacia un futuro más habitable, justo y respetuoso de los derechos de todas las personas.

JÓVENES + IA: PERSPECTIVAS DE UNA GENERACIÓN
JÓVENES + IA: PERSPECTIVAS DE UNA GENERACIÓN

IA Y DERECHOS HUMANOS

¿QUÉ DESAFIOS TENEMOS EN AMÉRICA LATINA?

En un momento en el que la inteligencia artificial y el uso masivo de datos están redefiniendo nuestras sociedades, resulta imprescindible abrir espacios de reflexión informada y crítica.

Con este espíritu, invitamos a personas especialistas, académicas y líderes involucradas en temas de datos, tecnología e inteligencia artificial a contribuir con artículos que exploran los desafíos éticos, sociales y políticos que acompañan a estos desarrollos. Su participación enriquece este libro con miradas diversas, rigurosas y profundamente conectadas con los debates contemporáneos sobre gobernanza digital, justicia algorítmica y derechos humanos. A continuación compartimos las reflexiones de **Nicole Angel Sánchez Rojas** de Fundación Internet Bolivia, **Mariel García Montes** de Massachusetts Institute of Technology, **Mario A. Sandoval M.** de ITESM-CCM/FFYL UNAM y **Violeta Belver** de la Iniciativa Latinoamericana por los Datos Abiertos. Este apartado reúne sus aportes con la convicción de que comprender la complejidad de la IA requiere escuchar voces expertas que, desde distintos países y sectores, analizan los riesgos y oportunidades de un futuro cada vez más automatizado.

Sesgos y discriminación en la inteligencia artificial, la censura digital en América Latina como dilema ético emergente

Autora: Magister Nicole Angel Sánchez Rojas, Fundación Internet Bolivia

La expansión de la inteligencia artificial (IA) y de los sistemas de moderación algorítmica ha transformado radicalmente los procesos de comunicación, protesta y participación política y ciudadana en el siglo XXI. Las plataformas digitales, redes sociales y motores de búsqueda aplican algoritmos para filtrar y priorizar contenidos, determinando quién los ve y quién no. En principio la tecnología busca dentro de sus principios principales la democratización del acceso a la información, además de multiplicar las voces en la esfera pública. Sin embargo, la realidad en varios países de América Latina, así como otros países parte del sur global, muestra un panorama más complejo, donde los algoritmos de moderación y las líneas editoriales en redes sociales producen sesgos estructurales como resultados de las opacas reglas que se tiene tras la moderación en plataformas digitales, que traen resultados como censura digital, shutdowns[1] (interrupción intencional de internet) o bloqueos de redes sociales, justificando las acciones en temas de seguridad ciudadana y estatal.

La situación internacional trae consigo cuestionantes, dentro de las cuales una de las principales es, **"¿cómo garantizar que la IA y las políticas de control digital respeten la equidad y la no discriminación cuando se aplican en sociedades ya desiguales?,** más aún en países con lineamientos raciales y discriminatorios específicos como ocurre en cada región y país del mundo.

En la región se observan líneas de desigualdad, fragilidad en las instituciones y violencia política, como obstáculos constantes a la libertad de expresión. Hoy vemos cómo todo esto aumenta los niveles de vulnerabilidad asociados al uso de IA y al control digital,

[1] Es una interrupción intencionada de las comunicaciones basadas en Internet, como el acceso a la web, datos móviles o servicios de mensajería, para controlar o silenciar a una población específica. Internet Society Position on Internet Shutdowns. (n.d.). Pulse.internetsociety.org. https://pulse.internetsociety.org/shutdown-statement

entre otros problemas, como la brecha en temas de alfabetización digital que crece con el avance de la tecnología.

En el año 2021 la UNESCO (UNESCO, 2021), recomendaba la consideración de contextos locales, buscando evitar que la tecnología se conviertiera en un mecanismo de exclusión. Este documento buscaba el desarrollo de una normativa de carácter no vinculante para la implementación de sistemas de IA con base en Derechos Humanos, por lo mismo, a diferencia de otros textos cuenta con un importante énfasis en la dignidad humana, justicia social, pluralidad de voces, gobernanza inclusiva, inclusión y diversidad cultural, equidad algorítmica y transparencia, prohibición de usos dañinos de la IA, puntos que deberían ser analizados en el desarrollo de políticas en temas de Inteligencia Artificial en América Latina, más allá del análisis de los principios de la OCDE o el AI Act.

En ese marco, este artículo explora cómo dichas recomendaciones cobran mayor relevancia cuando se las analiza desde los propios contextos latinoamericanos, caracterizados por desigualdades estructurales, fragilidad institucional y limitaciones en la capacidad regulatoria. La discriminación algorítmica y la censura digital no son fenómenos abstractos, estos se expresan en forma de shutdowns de redes sociales, en sesgos en la moderación de contenidos y en la propagación de desinformación electoral, problemáticas que reflejan tensiones internas de gobernanza democrática y brechas tecnológicas. A partir de estos ejemplos, se plantea una propuesta ética que prioriza la equidad, la transparencia y la protección de derechos humanos como principios rectores, formulando recomendaciones adaptadas a los desafíos y realidades de la región en la construcción de marcos normativos sobre Inteligencia Artificial.

Sesgos algorítmicos y discriminación en América Latina

Los sesgos algorítmicos no son un fenómeno accidental, son el resultado de la forma en la que se construyen y entrenan los sistemas de Inteligencia Artificial, esto se origina cuando los sistemas de IA son entrenados con datos históricos que contienen prejuicios, o cuando las decisiones de los ingenieros reflejan sus propias perspectivas. Investigaciones académicas han demostrado que los algoritmos pueden discriminar por raza, género o geografía, incluso sin un “intento” explícito. El aprendizaje automático depende de datos históricos que suelen reflejar desigualdades estructurales, lo que reproduce y amplifica discriminaciones preexistentes (Barocas & Selbst, 2016). Esto se puede observar en Latinoamérica de forma concreta y visible.

En 2021 en Colombia durante el desarrollo de manifestaciones pacíficas por parte de la población, se observaron acciones violentas por parte de efectivos policiales, sin embargo, el abuso no se limitó a las calles, donde las personas ejercían su libertad de expresión, ya que a través de redes sociales se vivía otro tipo de limitación de derechos como la libertad de expresión, los cuales se hacían evidentes tras la eliminación de transmisiones de abusos policiales y otras vulneraciones de derechos humanos. Según un reportaje de Meedan, muchos manifestantes denunciaron que sus publicaciones en Instagram, Twitter y Facebook eran eliminadas o invisibilizadas, la Fundación para la Libertad de Prensa advirtió que "estas prácticas constituían una violación de los derechos a la comunicación y a la libertad de expresión" (Meedan, 2024). Facebook respondió que sus sistemas de IA estaban entrenados para eliminar contenido violento y que tenían decenas de miles de moderadores; sin embargo, no aclaró cuántos de esos moderadores eran locales ni cómo sus algoritmos diferenciaban entre denuncias de abuso policial e incitación a la violencia

Un caso similar al colombiano se observó en Chile, el año 2019, donde videos, imágenes y post en formato texto denunciando abusos por parte de policías, fueron eliminados de redes sociales, y más aún en casos donde hashtags vinculados a denuncias sexuales eran parte de las publicaciones. Investigaciones impulsadas por el Observatorio del Derecho a la Comunicación y la Universidad de Chile registraron 283 incidentes de censura en redes sociales durante las protestas de 2019 (Del Campo, 2020). Un movimiento de protesta social denominado por algunas personas como el "femitag" un movimiento feminista que hacía visible no solo los abusos contra mujeres, sino también contra cualquier ciudadano y ciudadana en espacios de protestas. Según los reportes del Human Rights Watch del año 2020, en Chile se cometieron actos de violencia durante las detenciones realizadas por parte de policías, muchas personas indicaron ser víctimas de graves abusos durante su detención, estas denuncias mencionaban brutales golpizas y casos de abuso sexual. Los casos de censura incluyeron la eliminación de videos de movilizaciones y el cierre de cuentas de usuarios sin un mecanismo de apelación. Los investigadores señalaron que la automatización, la falta de contexto y la opacidad de las reglas de las plataformas fueron los principales causantes de estas decisiones

Estos movimientos de protesta en medios digitales es cada vez más constante para mostrar casos de abusos y vulneración de derechos humanos, pero al mismo tiempo se muestra como las redes sociales pueden ser espacios de vulneración de derechos como la libertad

de expresión y si bien, hasta hace unos años atrás estas moderaciones de contenidos tenían ciertas limitaciones, hoy en día la inteligencia artificial y los sesgos algorítmicos, así como las bases de datos utilizadas para entrenar a sistemas de Inteligencia Artificial muestran nuevos retos para la defensa de Derechos Humanos en entornos digitales.

Desde una perspectiva ética, estas situaciones vulneran el principio de no discriminación, al impactar de manera desproporcionada a comunidades históricamente marginadas. El sesgo no es sólo técnico, sino político, ya que decide qué voces son amplificadas y cuáles silenciadas.

Cuando se habla de idiomas o culturas originarias el sesgo y trato discriminatorio es aún más fuerte. Un estudio del Centro para la Democracia y Tecnología (Center for Democracy and Technology, 2025) sobre moderación en lengua quechua muestra que los usuarios de esta lengua enfrentan problemas recurrentes: publicaciones removidas sin explicación, algoritmos incapaces de diferenciar entre insultos y palabras indígenas y ausencia de moderadores que hablen la lengua. El informe concluye que los modelos de lenguaje de las plataformas no están preparados para idiomas de bajo recurso para el desarrollo de tecnología o plataformas web y que las políticas de moderación fueron copiadas del español o del inglés sin adaptarse al quechua. Esto evidencia una nueva forma de inequidad lingüística, donde las comunidades indígenas son marginadas del espacio digital por falta de representatividad en los datos de entrenamiento.

La discriminación algorítmica no se limita a la esfera social. En procesos electorales, la moderación sesgada puede influir en la opinión pública. En México, observatorios ciudadanos documentaron que durante las elecciones de 2018 y 2021 se etiquetaron publicaciones críticas como “fake news”, mientras que campañas oficiales con datos engañosos se mantuvieron visibles.

Conexiones entre la moderación algorítmica y control digital

Aunque la discriminación algorítmica y los apagones puedan parecer fenómenos distintos, en la práctica se entrelazan y se potencian. Las decisiones automatizadas de etiquetar un contenido como “sensitivo” o “incitador” sirven de argumento para que los Estados justifiquen bloqueos o regulaciones restrictivas. Paralelamente, el uso de apagones o bloqueos de redes acentúa la dependencia de las personas en plataformas que aplican reglas opacas, creando un círculo vicioso de control digital.

En Chile, el estudio sobre las protestas de 2019 encontró que muchos usuarios tuvieron sus cuentas suspendidas sin un recurso efectivo. La causa, según los autores, fue la automatización combinada con la falta de contexto, entre algoritmos entrenados con referentes de otros países no comprenden las expresiones locales y marcaban contenidos legítimos como infracciones y las líneas editoriales por parte del gobierno que invocó la presencia de "incitación a la violencia" en redes para presentar proyectos de ley que aumentarían la capacidad estatal de bloquear contenidos. Esta interacción demuestra cómo la moderación algorítmica puede ser instrumentalizada para legitimar la censura.

Durante las elecciones de Brasil 2022, la Corte Superior Electoral (TSE) firmó memorandos con Twitter, TikTok, Facebook y WhatsApp para frenar la desinformación. (Human Rights Watch, 2022). Sin embargo, organizaciones de la sociedad civil denunciaron que estas medidas fueron insuficientes: miles de publicaciones que alegaban fraude continuaron circulando, muchas de ellas impulsadas por políticos influyentes.

Por último, la clasificación automatizada de ciertos términos puede utilizarse como herramienta de persecución. Si los algoritmos de una red social asocian palabras como "territorio indígena" o "autonomía" con supuestas amenazas, esta información podría ser utilizada por las autoridades para justificar investigaciones o arrestos. En un contexto donde las leyes antiterrorismo son amplias, las etiquetas generadas por algoritmos pueden convertirse en pruebas incriminatorias, sin que exista transparencia sobre cómo se generaron.

Derechos que nos amparan y caminos de protección

A pesar del panorama preocupante que se desarrolla en líneas anteriores, no estamos del todo desprotegidos. Existen instrumentos internacionales que ofrecen un paraguas jurídico y ético para exigir que la tecnología respete nuestros derechos. En 2021, la UNESCO aprobó la Recomendación sobre la Ética de la Inteligencia Artificial, el primer acuerdo global de este tipo, si bien no es un documento vinculante para los estados, busca una aplicación ética y responsable por parte de los Estados. Este texto recuerda que la diversidad y la inclusión deben estar presentes en cada fase del desarrollo de sistemas de IA. También subraya que las tecnologías deben utilizarse de forma proporcional, evitando causar daños innecesarios.

En nuestra región, el Pacto de San José y la Declaración de Principios sobre Libertad de Expresión de la OEA son clave. El artículo 13 del Pacto garantiza el derecho de todas las personas a buscar, recibir y difundir información,

prohibiendo la censura previa y cualquier forma de control indirecto sobre los medios. (Mateo García Silva, & Maria Fernanda Chanduvi, 2024) Esto significa que los apagones de internet o los bloqueos masivos de redes sociales suelen ser incompatibles con la normativa regional salvo en casos muy excepcionales y deben pasar un test de legalidad, necesidad y proporcionalidad.

Mientras tanto, fuera de Latinoamérica, la Unión Europea se encuentra en el proceso de difusión para comenzar a implementar el AI Act, una propuesta que clasifica los sistemas de IA según su nivel de riesgo y exige obligaciones como la transparencia y las evaluaciones de impacto para los algoritmos de alto riesgo. Aunque no es aplicable aquí, puede servir de referencia para exigir reglas similares: auditar sistemas de moderación, informar sobre los criterios de eliminación de contenido y garantizar mecanismos de apelación.

A la luz de estos marcos, ¿qué medidas de protección se pueden aplicar? En primer lugar, exigir que las plataformas y los gobiernos respeten los principios de diversidad, inclusión y no discriminación. En segundo lugar, promover leyes nacionales que prohíban los apagones arbitrarios y obliguen a las plataformas a explicar sus decisiones. Y, finalmente, impulsar mecanismos de reparación efectivos: cuando un contenido se elimina injustamente o cuando un algoritmo discrimina, debe existir un canal claro para reclamar y obtener una solución.

Cuando la tecnología silencia, pero la sociedad resiste

Cuando algoritmos injustos se combinan con políticas de bloqueo digital, la democracia y los derechos humanos se ven directamente afectados. En lugar de integrar a todos, muchos sistemas de IA terminan amplificando la exclusión, ante esto la UNESCO advierte que estos sistemas deben promover la justicia social y evitar reforzar desigualdades, pero la ausencia de datos representativos y moderadores locales margina a mujeres, pueblos indígenas y comunidades LGBTIQ+. la opacidad de las plataformas: no desglosan sus datos ni explican sus decisiones, lo que impide que la ciudadanía evalúe el impacto de la moderación. Además, según Access Now, las manifestaciones siguen siendo uno de los principales detonantes de los apagones digitales, bloquear la red para detener protestas viola el derecho a la participación política y alimenta la impunidad.

La decisión del gobierno cubano de cortar internet y bloquear aplicaciones en 2021 demuestra cuán desproporcionado es dejar a toda una población sin acceso a la red. Este tipo de apagones no solo silencian denuncias: impiden acceder a servicios médicos, a la educación y a la

propia organización de protestas. De hecho, el informe de Access Now señala que las manifestaciones siguen siendo uno de los detonantes más comunes de los shutdowns. A esto se suma la opacidad de las plataformas: no desglosan sus datos por país o idioma y no explican por qué se elimina un contenido, lo que dificulta la rendición de cuentas. Y en las comunidades rurales y de bajos ingresos, la brecha digital se ensancha, pues dependen de un único proveedor y no tienen recursos para burlar la censura.

En el segundo semestre de 2025, muestra un nuevo caso, la reciente oleada de protestas en Nepal muestra hasta qué punto la combinación de censura digital y descontento social puede detonar crisis graves. A principios de septiembre de 2025 el gobierno nepalí prohibió 26 redes sociales y servicios de mensajería (incluidos Facebook, Instagram, WhatsApp, YouTube y X) argumentando que las empresas no se habían registrado bajo nuevas reglas. La medida provocó que decenas de miles de jóvenes salieran a las calles en la llamada “protesta Gen Z”; al intentar entrar en el parlamento, la policía respondió con munición real, gases lacrimógenos y balas de goma. El saldo fue de al menos 19 muertos y más de 200 heridos (Ellis Petersen, 2025) Aunque el gobierno levantó el veto, la indignación continuó y acabó con la dimisión del primer ministro, KP Sharma Oli. Este episodio demuestra que los apagones de internet, lejos de apaciguar, alimentan la protesta y conllevan graves violaciones de derechos humanos.

Frente a este panorama, hay salidas concretas. En primer lugar, exigir que los gobiernos renuncien a los apagones arbitrarios y que cualquier restricción cumpla con los principios de legalidad, necesidad y proporcionalidad que establecen la UNESCO y el Pacto de San José. También es clave auditar de forma independiente los algoritmos de moderación para detectar y corregir sesgos, y crear mecanismos de apelación claros para que los usuarios puedan reclamar y recibir reparación. La participación de comunidades marginadas en el diseño de la IA (incluyendo a hablantes de lenguas indígenas) es esencial para evitar inequidades. Asimismo, es necesario que las plataformas publiquen datos desglosados por país e idioma, para evaluar con precisión su impacto. Finalmente, una regulación regional armonizada, inspirada en la recomendación de la UNESCO y en los instrumentos interamericanos, unida a programas de educación digital, puede empoderar a la ciudadanía y convertir a América Latina en un referente de innovación tecnológica con justicia social.

Blindar nuestra voz digital en tiempos de algoritmos

El camino no es renunciar a la tecnología, sino exigir un marco que ponga a las personas y a sus derechos en el centro, conviene enfatizar que prohibir los apagones arbitrarios es un paso indispensable: cualquier restricción al acceso a internet debe respetar los principios de legalidad, necesidad y proporcionalidad que recogen la UNESCO y la Convención Americana.

Las plataformas tecnológicas, responsables de la moderación, deben someter sus algoritmos a auditorías independientes para detectar sesgos y crear canales de apelación claros y accesibles. Asimismo, la inclusión de pueblos indígenas y grupos marginados en el diseño de los sistemas de IA es esencial para evitar inequidades lingüísticas y culturales. Resulta prioritario exigir transparencia: las empresas deben publicar datos desglosados por país e idioma sobre qué contenidos retiran y por qué razones. A nivel político, la región necesita una agenda común que armonice las normas sobre inteligencia artificial y derechos digitales, inspirada en la recomendación de la UNESCO y en los instrumentos interamericanos, para que los Estados no actúen aislados frente al poder de las grandes plataformas. Finalmente, fortalecer la alfabetización digital de la ciudadanía permitirá comprender cómo operan los algoritmos y exigir rendición de cuentas. Solo así América Latina podrá construir un modelo de gobernanza digital que combine innovación tecnológica con justicia social y respeto a los derechos fundamentales.

Referencias

- Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104(3), 671–732. https://doi.org/10.2139/ssrn.2477899
- Censorship in Colombia's national strike #paronacional. (2024, February 5). Meedan.com. https://meedan.com/post/memory-and-censorship-in-colombias-national-strike-paronacional
- Center for Democracy and Technology (2025). Moderating Quechua Content on Social Media. https://cdt.org/insights/moderating-quechua-content-on-social-media/
- Del Campo, A. (2020), Moderate Globally Impact Locally: Social media in Latin America, Caught between a rock and a hard place. Yale Law School. https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-informati

on/wiii-blog/moderate-globally-impact-locally-social-media-latin-america-caught-between-rock-and-hard-place

- Ellis-Petersen, H. (2025). At least 19 killed in “gen Z” protests against Nepal’s social media ban. The Guardian; The Guardian. https://www.theguardian.com/world/2025/sep/08/nepal-bans-26-social-media-sites-including-x-whatsapp-and-youtube
- Informe Mundial (2020): Tendencias de los derechos en Chile. https://www.hrw.org/es/world-report/2020/country-chapters/chile
- Human Rights Watch (2022). Social Media Platforms Are Failing Brazil’s Voters. https://www.hrw.org/news/2022/10/28/social-media-platforms-are-failing-brazils-voters
- Mateo García Silva, & Maria Fernanda Chanduvi. (2024, July 8). A Review of Content Moderation Policies in Latin America. Tech Policy Press. https://www.techpolicy.press/a-review-of-content-moderation-policies-in-latin-america/
- UNESCO (2021). Recomendación sobre la ética de la inteligencia artificial https://www.unesco.org/es/articles/recomendacion-sobre-la-etica-de-la-inteligencia-artificial

Injusticia estructural y las decisiones automatizadas en México y América Latina

Autora: Mariel García Montes, Massachusetts Institute of Technology

Cuando escucho a alguien decir que la inteligencia artificial no forma parte de la vida de las personas en México a comparación de otras partes del mundo, pienso en mis encuentros diarios con sistemas donde se automatizó la toma de decisiones: banca, tiendas, algunos trámites gubernamentales. Sin embargo, mi creciente lista de experiencias diarias parecería estar en un universo paralelo, ajeno al universo de los espacios de deliberación sobre inteligencia artificial en América Latina, donde la discusión invariablemente gira en torno a escenarios hipotéticos para el futuro o de ejemplos pasados que vienen de otras regiones del mundo.

En esas discusiones, es común que se haga referencia a casos e investigación del norte global. Es cierto que son casos emblemáticos, ampliamente documentados en medios de comunicación electrónicos internacionales. Estos casos se han prestado para ilustrar las preocupaciones de lo que podría suceder en potencia en América Latina si se llegaran a importar las tecnologías en cuestión. En lo que concierne a las decisiones automatizadas, generalmente se habla del papel de la inteligencia artificial en los siguientes casos:

1) *COMPAS*, un *software* de manejo de casos de personas privadas de la libertad en Estados Unidos, el cual está basado en un algoritmo de predicción de reincidencia criminal que busca guiar a los jueces en su toma de decisiones sobre fianzas. Este caso llegó a la discusión pública gracias al trabajo de la periodista de investigación Julia Angwin en ProPublica, cuyo equipo analizó los casos de 10,000 personas acusadas de crímenes, concluyendo que el algoritmo llevaba dos veces más a predicciones erróneas para las personas negras que para las personas blancas.[2]

Se trata de un ejemplo célebre en

[2] Jeff Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm", *ProPublica*, 23 de mayo de 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

las discusiones sobre decisiones automatizadas porque estudia el uso de un algoritmo predictivo para determinar si alguien debe obtener fianza, ir a prisión o recibir otros castigos según su supuesto riesgo de incidencia. Este caso ejemplifica una iniciativa tecnológica para la toma de decisiones que terminó por reforzar el racismo y la injusticia estructural.

2) La experiencia en Holanda del uso de algoritmos anti-fraude por parte de la autoridad tributaria, especialmente en el padrón de beneficiarios del programa social para infancias. En este caso, la Administración de Impuestos y Aduanas holandesa automatizó sus decisiones sobre casos de fraude durante catorce años, lo que llevó a la acusación de 26,000 padres y el ingreso de más de mil infantes al sistema de acogida estatal. Eventualmente, una auditoría reveló que algunos de los criterios que la Administración eligió para determinar el riesgo de fraude eran la doble nacionalidad, el nivel socioeconómico bajo y la apariencia "no occidental".[3]

Este caso ha sido popular en las discusiones sobre decisiones automatizadas porque ejemplifica los peores daños que puede generar el uso, sin supervisión, de este tipo de algoritmos en la función pública. Los perjuicios que tuvo esta serie de errores, en las vidas de decenas de miles de familias y de mil infantes separados de sus padres, son irreparables.

Los ejemplos anteriores, producto de la investigación que surge gracias a la conjunción de periodismo, sociedad civil organizada y academia en el norte global, tienen utilidad discursiva para los esfuerzos en nuestra región. Es posible comprender, a partir de las experiencias de otros, cómo es que los avances de inteligencia artificial han contribuido a la injusticia estructural. Sin embargo, centrar estos ejemplos en nuestros diálogos en pos de regulación en América Latina nos impide responder al llamado de ir más allá de la tropicalización de los conceptos del norte, e ir hacia la teorización y comprensión de y desde las realidades de la mayoría global.[4]

[3] Politico, "Dutch scandal serves as a warning for Europe over risks of using algorithms," *Politico*, 29 de marzo de 2022, https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

[4] Edgar Gómez-Cruz et al., "Beyond the Tropicalization of Concepts: Theorizing Digital Realities with and from the Global South (Introduction to a Special Issue)," *Communication, Culture & Critique* 16, no. 4 (2023): 217–20, https://doi.org/10.1093/ccc/tcad037.

La automatización de las decisiones ya se da en América Latina. Más allá de tropicalizar la resistencia del norte global, necesitamos comprender y atender de manera amplia los orígenes, operaciones y consecuencias de los sistemas de decisión automatizada en nuestra región. Sólo así podremos enfrentar la prominencia de la inteligencia artificial en la vida pública. Es por eso que, en este ensayo, propongo una pregunta guía para las discusiones sobre la ética de las decisiones automatizadas en América Latina y del uso, en general, de herramientas computacionales para apoyar o sustituir la toma de decisiones: ¿Cómo es que los sistemas computacionales que influyen en la toma de decisiones podrían, a pesar de sus mejores intenciones, reforzar la desigualdad e injusticia estructurales?

Injusticia estructural y decisiones automatizadas en México y América Latina

Las amplias desigualdades de América Latina hacen que se convierta en el territorio perfecto para lo que Ricaurte, Gómez-Cruz y Siles llaman *gubernamentalidad algorítmica*: un poder neocolonial que permite la extracción que las compañías hacen de los recursos en América Latina y, a los gobiernos, automatizar las asimetrías sociales y el control social.[5] Este poder se ejerce cuando en América Latina el gobierno y las entidades privadas emplean las llamadas "soluciones algorítmicas" con el afán de responder a problemas sociales mediante el uso de tecnología. Ricaurte *et al*. nos recuerdan que estas decisiones tecno-determinísticas "legitimizan regímenes autoritarios y discriminatorios al reforzar la noción que ciertas personas (migrantes, jóvenes, mujeres y personas de ingresos bajos) representan un riesgo mayor".[6]

El trabajo conceptual de Ricaurte *et al*. forma parte de un cuerpo cada vez más grande de investigación alrededor del mundo que busca enunciar las maneras en que los proyectos tecnopolíticos basados en inteligencia artificial refuerzan las desigualdades alrededor del mundo. Este cuerpo ha logrado movilizar marcos conceptuales como *colonialismo de datos*, *feminismo de datos,* así como conceptos más específicos como el sesgo y los daños algorítmicos representacionales.[7] A mi

[5] Paola Ricaurte et al., "Algorithmic Governmentality in Latin America: Sociotechnical Imaginaries, Neocolonial Soft Power, and Authoritarianism," *Big Data & Society* 11, no. 1 (2024): 20539517241229697, https://doi.org/10.1177/20539517241229697.
[6] Paola Ricaurte et al., *ibid*, p. 4.
[7] Cf. el trabajo de Nick Couldry y Ulises Mejías, "Data Grab: The New Colonialism of Big Tech and How to Fight Back"; Catherine D'Ignazio y Lauren Klein, "Data Feminism"; Joy Buolamwini y Timnit Gebru "Gender shades: Intersectional accuracy disparities in commercial gender classification"; y Tarleton Gillespie "Generative AI and the politics of visibility".

parecer, este campo se puede también enriquecer con el análisis de la injusticia estructural.

La injusticia estructural se da cuando las prácticas sociales en la vida diaria ponen, de manera sistemática, a algunas personas en posiciones de mayor vulnerabilidad y a otras en posiciones de mayor seguridad. Aunque el concepto se basa en la teoría de la justicia de John Rawls, la politóloga Iris Marion Young desarrolló la teoría enfocada específicamente en las estructuras sociales que sostienen la injusticia estructural. Estas estructuras son las normas materiales, legales y sociales que determinan el rango de opciones que las personas tienen a su disposición y que resultan de decisiones pasadas.

Iris Marion Young propone el concepto, entonces, de injusticia estructural, definiéndola como un fenómeno que existe cuando "los procesos sociales ponen a grandes grupos de personas bajo la amenaza sistemática de dominación o de privación de los medios para desarrollar y ejercer sus capacidades, al mismo tiempo que estos procesos permiten a otros que dominen o tengan un amplio rango de oportunidades para desarrollar y ejercer las capacidades que tienen disponibles".[8] Young explica la injusticia estructural con el caso de *Sandy*, una madre soltera que se encuentra al borde de la indigencia, cuando una constructora compra el edificio donde ella alquila, y ella no puede encontrar otro lugar donde vivir.

En el caso de Sandy, la catástrofe es una confluencia de las acciones de muchas personas e instituciones que nunca buscaron despojarla, pero tuvieron ese efecto en agregado, en la estructura social que determinaron para ella. Trabajo posterior extiende el caso de Sandy a las discusiones de responsabilidad algorítmica: *Mandy* es una versión de Sandy que va al banco y, gracias a un algoritmo informado por valores de justicia distributiva, da el mayor crédito posible a Mandy para que pueda alquilar un lugar distinto. Sin embargo, termina sólo por reforzar la injusticia estructural que la rodea: la vivienda de Mandy peligra dado que ella ahora tiene una montaña de deudas.[9]

La injusticia estructural, así como la gubernamentalidad algorítmica, son una lente para articular que las buenas intenciones, individuales o institucionales, no necesariamente llevan a resultados buenos. En manos del gobierno y de

[8] Iris Marion Young, "Structure as the Subject of Justice," *Responsibility for Justice*, Oxford Political Philosophy (Oxford University Press, 2011).

[9] Atoosa Kasirzadeh, "Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy," *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, July 26, 2022, 349–56, https://doi.org/10.1145/3514094.3534188.

entidades privadas, los sistemas computacionales que influyen en la toma de decisiones, inclusive los diseñados con buenas intenciones, están encaminados a reforzar la desigualdad en la región. La pregunta central de este ensayo es *cómo*. Para responderla, analizo un caso de toma de decisiones automatizadas actualmente en desarrollo en México: la CURP biométrica, una propuesta de uso de la identidad biométrica para la automatización de la burocracia en México.

Uso de la identidad biométrica para automatizar la burocracia en México

En 2025, el Congreso de México votó a favor la instauración de un documento nacional de identificación conocido como *CURP biométrica*. Con este documento, el gobierno busca satisfacer su obligación, marcada en la Ley General de Población, de otorgar un documento de identificación a todas las personas mexicanas. De manera previa a la CURP biométrica, y desde 1992, el principal documento de identificación biométrica en México ha sido la credencial de votar con fotografía y huella dactilar, emitida por el Instituto Nacional Electoral. Para obtener dicha credencial, cada ciudadano se debe inscribir al cumplir 18 años en el Registro Federal Electoral. En las tres décadas de su validez como identificación oficial, el proyecto ha implicado un alto gasto para el país; tan sólo en la campaña de 1992-1993, México invirtió 2 billones de pesos. En 2025, destina mil 438 millones de pesos para el programa.[10] Sin embargo, la inversión ha tenido éxito: la adopción de este documento fue tal que, en 2024, el Instituto celebró la inscripción de 100 millones de personas mexicanas a este registro.[11]

Si la tasa de adopción de la credencial para votar es tan alta en México, ¿qué razones tendría el gobierno mexicano para instaurar un nuevo documento de identificación biométrica? En primer lugar, cita su obligación legal establecida en la Ley General de Población. La credencial de votar no es expedida por el gobierno, sino por un órgano autónomo que surgió en un momento de crisis electoral; su carácter como documento de identificación oficial se había planteado como temporal. Por otra parte, el proyecto dice que incorporará a una población que el Registro Federal de Votantes no puede: las personas mexicanas menores de 18 años.

Sin embargo, la motivación gubernamental principal para la CURP

[10] Ángel Cabrera, "Invierte INE 1,438 mdp en biométricos de ciudadanos", 24 Horas, 15 de julio de 2025, https://24-horas.mx/mexico/invierte-ine-1438-mdp-en-biometricos-de-los-ciudadanos/.

[11] Rosalía Vergara, "Ya hay más de 100 millones de mexicanos en el padrón electoral: INE", *Proceso*, 23 de enero de 2024. https://www.proceso.com.mx/nacional/2024/1/23/ya-hay-mas-de-100-millones-de-mexicanos-en-el-padron-electoral-ine-322639.html

biométrica, y donde también se maximizan los riesgos de injusticia estructural y gubernamentalidad algorítmica, es la custodia de los datos biométricos. Por ley, al estar el Registro Federal Electoral bajo resguardo del órgano electoral autónomo, los gobiernos en turno no han tenido acceso a la base de datos biométricos resultante. Esto significa que todas las adquisiciones de tecnologías de reconocimiento facial, como las cámaras de seguridad pública, no han podido usar el *corpus* de datos de entrenamiento algorítmico más grande en el país. Es decir: el registro biométrico hasta ahora más importante en el país, por sus protecciones constitucionales, no ha servido para propósitos de vigilancia masiva. Eso es lo que busca cambiar la CURP biométrica.

La CURP biométrica consiste en la recolección de tres datos biométricos para su resguardo en una Plataforma Única de Identidad: el iris, la fotografía y la huella digital, vinculadas a datos textuales como la clave única.[12] Aunque solamente aparecen estos datos biométricos en los documentos oficiales, distintos medios de comunicación también mencionan la grabación de voz, la cual el gobierno de México ya recolecta para los servicios de pensiones. Este proyecto es, entonces, una aplicación de la identificación computarizada de la biometría, tales como el reconocimiento facial y el análisis de huellas dactilares. Se trata de una forma de visión computarizada que abstrae los rasgos humanos a polígonos, objetos numéricos, permitiendo así su análisis masivo en grandes bases de datos. A nivel mundial, es una de las formas de inteligencia artificial que más han habilitado el patrullaje predictivo y la toma de decisiones automatizada, a pesar de las críticas por sus sesgos raciales y sus varias formas de discriminación.[13] Sin los controles necesarios, que por circunstancias históricas los datos biométricos del Instituto Nacional Electoral sí tienen, se vuelve una herramienta de control social.

No es la primera vez que surge esta aspiración, ya que la CURP biométrica no es el primer proyecto gubernamental para establecer un documento (y, especialmente, un registro) biométrico por el gobierno en turno. La inspiración para estas iniciativas viene de las cédulas de identidad de otros países en América

[12] Secretaría de Gobernación. "Decreto por el que se reforman, adicionan y derogan diversas disposiciones de la Ley General en Materia de Desaparición Forzada de Personas, Desaparición Cometida por Particulares y del Sistema Nacional de Búsqueda de Personas, así como de la Ley General de Población, en materia de fortalecimiento de búsqueda, localización e identificación de personas desaparecidas." *Diario Oficial de la Federación*, 16 de julio de 2025. https://dof.gob.mx/nota_detalle.php?codigo=5763157&fecha=16/07/2025#gsc.tab=0

[13] Joy Buolamwini y Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". *Conference on Fairness, Accountability and Transparency*. PMLR, 2018.

Latina, como el DNI que en Argentina emite el Registro Nacional de las Personas, y la Cédula de Identidad asociada al Rol Único Nacional en Chile. Desde el sexenio del Presidente Vicente Fox (2000-2006), cada administración ha intentado establecer su propio documento a través del Registro Nacional de la Población, pero ningún proyecto ha sido completado por falta de presupuesto, retos técnicos, y por voluntad política.

En 2025 surge nuevamente un proyecto alentado por dos promesas tecnopolíticas. En primer lugar, la Presidenta Claudia Sheinbaum la vincula con una de las crisis de derechos humanos más importantes que enfrenta México: la desaparición forzada. La Presidenta promete que la CURP biométrica será un mecanismo imprescindible para la búsqueda de las más de 130,000 personas desaparecidas en el país.[14] En segundo lugar, esta nueva forma de identificación permitirá a los ciudadanos facilitar los procesos burocráticos necesarios en la vida diaria en "bancos, hospitales, museos y para realizar distintos trámites".[15]

¿Cómo es que la CURP biométrica sirve como un mecanismo de decisión automatizada que refuerza la injusticia estructural en el país? De tres maneras: reemplazando una base de datos biométricos que es segura, permitiendo la recolección de datos sin controles que eviten la vigilancia masiva, y haciendo una promesa tecnopolítica falsa sobre el poder de la cédula de identidad biométrica para la localización de las personas desaparecidas en México.

1. Reemplazo de una forma de identificación ampliamente adoptada

En una conferencia de prensa en julio de 2025, la presidenta de la Comisión de Gobernación del Senado de la República, Margarita Valdez, recibió preguntas sobre la CURP biométrica y el intercambio de datos del gobierno mexicano con el de Estados Unidos. En sus declaraciones, confirmó el propósito de eliminar el carácter oficial como documento de identificación de la credencial de elector. "En el momento en que esté todo regularizado, nada más va a ser un documento para votar", declaró, y

[14] Gobierno de México, *Versión estenográfica. Conferencia de prensa de la presidenta Claudia Sheinbaum Pardo del 24 de julio de 2025*. https://www.gob.mx/presidencia/articulos/version-estenografica-conferencia-de-prensa-de-la-presidenta-claudia-sheinbaum-pardo-del-24-de-julio-de-2025

[15] Redacción El Financiero, "¿Adiós a la credencial del INE como identificación oficial? Este documento te pedirán en bancos y trámites", *El Financiero*, 10 de julio de 2025. https://www.elfinanciero.com.mx/nacional/2025/07/10/adios-a-la-credencial-del-ine-como-identificacion-oficial-este-documento-te-pediran-en-bancos-y-tramites/

también que "tomará tiempo que la gente se acostumbre al cambio".[16]

El desplazamiento del documento del INE tiene consecuencias estructurales. En primer lugar, al quitar la doble función de la credencial de votar, el único incentivo para obtener esta credencial es el derecho al voto. Consejeros del INE advierten que esta medida puede tener efectos negativos en el esfuerzo nacional de registro de votantes, con el potencial de que se incremente el abstencionismo.[17] Con esta iniciativa, no sólo será más difícil registrar a nuevas personas, sino también mantener registradas a las personas que ya lo están actualmente, pues la credencial tiene una vigencia de 10 años. Asimismo, cabe preguntarse qué efecto tendrá particularmente en las comunidades más remotas del país; si la inversión en la CURP biométrica será suficiente para alcanzarles desde el inicio, o si *de facto* quedarán desconectadas de los servicios comunidades enteras por la conjunción de su posición geográfica y la falta de identificaciones oficiales.

[16] Luis Enrique Orduna Ramirez, "¿CURP Biométrica sustituirá al INE como identificación oficial? Esto sabemos", *24 Horas*, 10 de julio de 2025. https://24-horas.mx/mexico/curp-biometrica-sustituira-al-ine-como-identificacion-oficial-esto-sabemos/#google_vignette

[17] Yared de la Rosa, "CURP biométrica desplazará la credencial de elector como identificación oficial", *Expansión Política*, 9 de julio de 2025. https://politica.expansion.mx/mexico/2025/07/09/curp-biometrica-obligatoria-desplazara-ine

2. Permite la recolección de datos para la vigilancia masiva, sin controles

En lo que concierne el desplazamiento del Registro Federal Electoral, la iniciativa de CURP biométrico limita y sustituye los billones de pesos de inversión que México ha hecho en una base de datos robusta, ampliamente adoptada, con usos no electorales bien delimitados y con salvaguardas de protección que han funcionado en los más de 30 años de su existencia. A lo largo de su existencia, el INE demostró su capacidad de gestionar el registro con protecciones para la ciudadanía y, al mismo tiempo, colaborar con jueces en casos individuales, así como con otras instituciones de maneras que protegieran los datos biométricos de los mexicanos en su padrón.[18] Las colaboraciones se dan mediante convenios, celebrados con instituciones bancarias y crediticias, en los que el INE valida en tiempo real la identificación de una persona dando un algoritmo de probabilidad.

En su lugar, el gobierno de la Presidenta Claudia Sheinbaum propone la creación de un registro completamente nuevo, que requiere que los 100 millones de mexicanos ya registrados para votar se tengan que acercar nuevamente al gobierno a dar aún más datos

[18] Alberto Alonso y Coria, en entrevista con la autora, 11 de marzo de 2024.

biométricos que antes y sin poder ofrecer, por la falta de controles institucionales, el mismo cuidado de los datos ante la realidad de vigilancia masiva que se ha documentado en la última década en el país.[19]

3. Promesa tecnopolítica falsa para las familias de personas desaparecidas

En una de las conferencias de prensa sobre la CURP biométrica, la senadora Reyna Celeste Ascencio mencionó que ésta surgió de las mesas de trabajo entre la Secretaría de Gobernación, familias de desaparecidos y asociaciones civiles:

> Lo que nosotros hemos dicho, a ver, si desaparece una persona y no damos con su paradero en un lapso de tantas horas, porque no sabemos cómo poder vincular rápidamente, pues a través de este nuevo instrumento. Es una cuestión que si lo vemos es herramientas digitales que nos permitan ser un país de avanzada.[20]

La promesa tecnopolítica más importante que subyace a la CURP biométrica es que permitirá la localización con vida de las personas que desaparecen en México. Sin embargo, los análisis de la sociedad civil organizada y la academia especializada en datos de la desaparición en México muestran que difícilmente este ejercicio biométrico sería suficiente para resolver la profunda problemática de la localización e identificación de las personas desaparecidas. En primer lugar, el proyecto de CURP biométrica como mecanismo de localización de personas desaparecidas no resuelve los siguientes retos que ha indicado la sociedad civil:

1) La falta de periodicidad consistente en la publicación de datos por parte de la Comisión Nacional de Búsqueda, y la falta del cumplimiento del requisito en la Ley General en Materia de Desaparición Forzada, Desaparición cometida por Particulares y del Sistema Nacional de Búsqueda de Personas de un

[19] Ver la crítica de la Red en Defensa de los Derechos Digitales: "La CURP biométrica obligatoria y la Plataforma Única de Identidad consolidarían un sistema autoritario de vigilancia masiva en México", 3 de abril de 2025, https://r3d.mx/2025/04/03/la-curp-biometrica-obligatoria-y-la-plataforma-unica-de-identidad-consolidarian-un-sistema-autoritario-de-vigilancia-masiva-en-mexico/.

[20] Grupo Parlamentario Morena, LXVI Legislatura. "Conferencia de prensa de las senadoras Lilia Margarita Valdez y Reyna Celeste Ascencio. Versión estenográfica", *Senadores Morena LXVI Legislatura*, 9 de julio de 2025, https://morena.senado.gob.mx/conferencia-de-prensa-de-las-senadoras-lilia-margarita-valdez-y-reyna-celeste-ascencio/.

portal de datos abiertos sobre los casos de desaparición;[21]

2) las omisiones dentro de los procesos de registro de personas desaparecidas, donde, por ejemplo, menos del 2% de los casos cuentan con el registro tan sólo de un medio de contacto con el ser querido que levanta el reporte;[22]
3) la falta de garantía del funcionamiento de varios registros: el Banco Nacional de Datos Forenses, el Registro de Fosas Comunes y Fosas Clandestinas, el Registro Nacional de Personas Fallecidas No Identificadas y no Reclamadas;[23]
4) la falta de interoperabilidad entre las fiscalías estatales y General, las comisiones estatales y la Nacional de Búsqueda, el Servicio Mexicano Forense y los sistemas de morgues;[24]
5) las discrepancias entre los datos oficiales y los datos de familiares y de personas que investigan la desaparición en México;[25]
6) la trayectoria de violaciones de los datos biométricos de desaparecidos y de sus familiares en un país donde las fiscalías han dado a empresas privadas acceso a datos genéticos para comercializar el dolor de las personas que buscan a sus seres queridos.[26]

Pensar la CURP biométrica como solución antes de atender los problemas de las plataformas actuales, cuyo funcionamiento tendría que ser óptimo para que la CURP biométrica pudiera asistir en las investigaciones, afecta de manera desproporcionada a las personas que ya cargan los peores daños de la tarea de buscar a sus seres queridos.

Conclusión

Las discusiones sobre regulación de las decisiones automatizadas en América

[21] Data Cívica, "A quienes nos faltan: datos para encontrarles", *Data Cívica*, agosto de 2024, p. https://media.datacivica.org/pdf/aquienesnosfaltan-2024-DATACIVICA.pdf, 21-22.
[22] Ibid., 16.
[23] Ver el amparo de Olimpia Montoya, representada por el Centro ProDH, contra la Fiscalía General de la República por su omisión de crear el Banco Nacional de Datos Forenses, así como los resultados de la solicitud de transparencia de sus abogados, que demuestra la discrepancia de perfiles genéticos entre las fiscalías y la FGR. Efraín Tzuc, "Juez Federal resolverá demanda contra FGR por omisión de crear el Banco Nacional de Datos Forenses", *A dónde van los desaparecidos*, 4 de abril de 2022, https://adondevanlosdesaparecidos.org/2022/04/04/juez-federal-resolvera-demanda-contra-fgr-por-omision-de-crear-el-banco-nacional-de-datos-forenses/; Efraín Tzuc, "¿Qué perdemos con la desaparición del INAI?", *Gatopardo*, 14 de noviembre de 2024, https://www.gatopardo.com/articulos/que-perdemos-con-la-desaparicion-del-inai.
[24] Ibid.
[25] Data Cívica, *Volver a desaparecer*, consultado en agosto de 2025. https://volveradesaparecer.datacivica.org/
[26] Paula Mónaco Felipe y Wendy Selene Pérez, *Traficantes de ADN*, 13 de diciembre de 2021. https://traficantesdeadn.com/

Latina deben alejarse de los casos hipotéticos y de los casos internacionales para también centrarse en los daños que ya trae esta práctica a nuestros contextos. Estos daños, entendidos a través de lentes como la teoría de la injusticia estructural, pero también lentes latinoamericanos como la gobernabilidad algorítmica, articulan cómo las buenas intenciones pueden concentrar sus daños en los grupos minorizados. En el caso de México, uno de los ejemplos más apremiantes es la CURP biométrica, que propone sustituir la base de datos biométricos de más de 100 millones de mexicanos por una base completamente nueva. Es decir, cambiar de una base que hasta ahora ha tenido uso restringido y autónomo, a una nueva sin los controles institucionales necesarios para garantizar el respeto a los derechos de las personas en México. Este proyecto, que es uno necesario para los procesos de automatización de decisiones gubernamentales, se hace en nombre de las personas que han sufrido los peores daños de la desaparición forzada en el país, haciéndoles una promesa tecnopolítica falsa sobre el poder de una base de datos para localizar a sus seres queridos con vida.

La sociedad civil como pilar de la gobernanza de la IA en América Latina: transparencia, rendición de cuentas y desempeño democrático

Autora: Violeta Belver, ILDA

En América Latina, la adopción de sistemas de inteligencia artificial (IA) se acelera con un impulso destacado en áreas clave como servicios públicos, industrias estratégicas y la gestión estatal. La promesa de eficiencia y modernización que conlleva la IA coexiste, sin embargo, con una realidad institucional que enfrenta dificultades para alinear la tecnología a principios de derechos humanos. Los hallazgos del Índice Global de IA Responsable (GIRAI) confirman la brecha: sociedad civil y academia protagonizan, de manera sistemática, la agenda de IA responsable en contextos donde los Estados demoran en diseñar y aplicar normativas exigibles, métricas de desempeño y mecanismos de rendición de cuentas. Este ensayo parte de esa constatación para argumentar que el rol de la sociedad civil no es suplementario, sino estructural, y que su contribución más significativa consiste en traducir principios en prácticas verificables que eleven el estándar de gobernanza democrática de la IA.

Contexto en América Latina

El GIRAI (2024) identifica, de manera transversal, dos tendencias relevantes para América Latina. Primero, una significativa demora por parte de los gobiernos para traducir principios a mecanismos exigibles en áreas críticas como evaluaciones de impacto, transparencia y explicabilidad, participación significativa y reparación. Segundo, el liderazgo de sociedad civil y academia en generación de evidencia, vigilancia pública y diseño de estándares prácticos. Esta preeminencia de acciones por parte de sectores no estatales se destaca especialmente en dimensiones asociadas a derechos: allí donde faltan evaluaciones de impacto algorítmico o rutas claras de reparación, los avances provienen de investigación aplicada, observatorios y litigio estratégico que vuelven visibles daños y vacíos institucionales.

A nivel internacional, podemos identificar marcos de referencia como la Recomendación sobre la Ética de la

Inteligencia Artificial de UNESCO (2021), que reconoce que la gobernanza de la IA requiere dispositivos adaptativos y colaboración multiactor a lo largo de todo el ciclo de vida de los sistemas, vinculando transparencia, explicabilidad, responsabilidad y reparación. La distancia entre ese estándar y la práctica regional se explica, en parte, por la falta de instrumentos estabilizados de transparencia que habiliten monitoreo ciudadano y rendición de cuentas.

Gobernanza rezagada y asimetrías de poder técnico

El desfase entre innovación y regulación, presente en numerosas regiones, se agrava en América Latina por tres factores que interactúan y se refuerzan. Primero, la dependencia tecnológica y contractual respecto de proveedores que operan bajo regímenes jurídicos y culturales distintos, con cláusulas de secreto comercial que bloquean la auditoría independiente. Segundo, déficits de capacidad institucional que dificultan transformar principios en dispositivos operativos: catálogos de sistemas en uso, reglas de ciclo de vida de datos, criterios de proporcionalidad, explicación de decisiones con efectos materiales y rutas claras de reparación. Tercero, la ausencia de espacios de prueba y evaluación *ex ante* para sistemas de alto riesgo, lo que traslada la experimentación al entorno real y desplaza los costos del error a poblaciones vulnerables. El GIRAI también revela que, donde estas condiciones se observan, sociedad civil y academia asumen el liderazgo en la generación de evidencia, la vigilancia pública y la formulación de estándares para orientar compras, evaluaciones y auditorías.

Transparencia algorítmica: avances, límites y madurez pendiente

Hablar de transparencia algorítmica hoy en América Latina es describir una zona de experimentación con herramientas y formatos cuya madurez está aún en construcción. Dos reportes recientes ilustran esa trayectoria:
El informe “Transparencia algorítmica en el sector público: estado del arte de instrumentos” (GPAI/OCDE, 2025) sistematiza la noción de transparencia algorítmica como capacidad, principio, norma/regla e instrumento, y analiza 83 repositorios públicos a escala global, con tipologías de instrumentos basados en la oferta y la demanda.

Por otro lado, “Hacia una transparencia algorítmica” (2025), reporte publicado por Artículo 19 y la Thomson Reuters Foundation, mapea repositorios y registros en el mundo y en la región, ofreciendo un catálogo de variables mínimas, organizadas por etapas del ciclo de vida del sistema. La conclusión converge con la del informe del GPAI: hay registros, guías y protocolos, pero predomina una transparencia mínima

que publica existencia, proveedor y propósito, sin documentación técnica, métricas ni resultados de auditorías que habiliten evaluación externa del desempeño.

Ambos diagnósticos identifican una tensión de fondo: tecnologías intensivas en datos y modelos complejos demandan herramientas de gobernanza que, a menudo, no existen o no se aplican. En la práctica regional emergen tres fuentes de opacidad:

- Opacidad contractual y secreto comercial. Las compras públicas ceden información técnica crítica a proveedores, limitando auditorías externas. El informe del GPAI subraya que, sin cláusulas de transparencia y acceso a interfaces/logs, la auditabilidad es inviable.

- Falta de catálogos actualizados de sistemas. Sin registros públicos que indiquen qué sistemas están en uso, en qué dominios, con qué datos y bajo qué salvaguardas, la rendición de cuentas se vuelve imposible de concretar.

- Complejidad técnica y datasets no públicos. Algunos modelos, por diseño y entrenamiento, resisten la inspección externa, por lo que si no existen regímenes de acceso condicionado para auditorías, la transparencia queda limitada a fichas informativas.

Frente a este cuadro, la pregunta por la transparencia se vuelve más importante que nunca. Una transparencia que no permite comprensión ni control no va lograr un impacto real para avanzar a sistemas de IA que estén alineados a los principios de derechos humanos y aporten al bien común. Necesitamos prácticas de transparencia algorítmica que incluyan, consideren y habiliten a distintos públicos a comprender lo relevante para ejercer sus derechos, indagar justificaciones y activar mecanismos de reparación. Sin dudas, esto requiere decisiones de política pública, capacidades técnicas y acompañamiento externo de actores de sociedad civil y academia.

Sociedad civil para avanzar a una IA responsable

En la discusión sobre una gobernanza responsable y ética de la IA, la sociedad civil desempeña un doble papel. Por un lado, produce conocimiento allí donde el Estado no llega: mapea algoritmos, documenta sesgos y efectos, crea catálogos cívicos y repositorios que permiten seguir el rastro de sistemas y decisiones. Por otro, impulsa exigencias para que ese conocimiento se incorpore en la toma de decisiones públicas: desde cláusulas contractuales que exigen

documentación técnica esencial, hasta la promoción de registros que incorporan variables de ciclo de vida, como datos, métricas, frecuencia de actualización, y la demanda de *sandboxes* regulatorios para someter a prueba sistemas de alto riesgo antes de su despliegue.

La contribución más influyente de la sociedad civil no se agota en el señalamiento de problemas: su valor también está en el diseño de formatos de respuesta que transforman estándares difusos en reglas operativas. Modelos de declaración de transparencia que interrogan por calidad de datos, controles humanos y evaluaciones de impacto; guías de compras públicas que incorporan criterios de proporcionalidad, acceso a logs y mecanismos de auditoría; propuestas de evaluación de impacto algorítmico que combinan análisis técnico con evaluación de riesgos a derechos y contextos de uso. En varios casos, estas propuestas culminan en acuerdos formales que condicionan adquisiciones o en decisiones judiciales que obligan a abrir información. Más allá de sus resultados inmediatos, desplazan el umbral de legitimidad: lo que ayer se consideraba secreto comercial intocable hoy puede, razonablemente, tratarse como información de interés público cuando impacta derechos.

Participación significativa y capacidades

La participación pública en proyectos de IA aparece en la mayoría de marcos internacionales o regionales como un principio esencial, pero en la realidad, rara vez se operacionaliza con efectos verificables.

Una participación significativa no es un taller perdido en el ciclo de desarrollo o implementación de la IA o una consulta simbólica, sino la incorporación de comunidades afectadas en la definición del problema, el diseño de soluciones, la evaluación pre-despliegue y el monitoreo post-despliegue. Esa inclusión no solo mejora legitimidad o modifica decisiones técnicas, sino que redefine variables relevantes, revela datos faltantes y anticipa contextos de uso que los equipos de desarrollo pueden no identificar.

A su vez, ninguna arquitectura de gobernanza funciona sin capacidades institucionales. Formación en evaluación de impacto, gestión de riesgos, contratación responsable, seguridad de la información y gobernanza de datos no son complementos, sino condiciones necesarias para que las reglas operen. En este sentido, la sociedad civil ha cumplido un papel pedagógico decisivo, formando funcionarios, periodistas y organizaciones locales, y generando materiales abiertos y relevantes que aumentan la agencia ciudadana para monitorear.

Equidad, interseccionalidad y sostenibilidad

La promesa de la IA en políticas públicas se juega, en gran medida, en su capacidad para reducir desigualdades. Para constatar y evaluar avances respecto de ese objetivo, es fundamental contar con métricas de impacto desagregadas por variables relevantes (género, raza, pueblos indígenas, discapacidad, territorio, entre otras), así como con coproducción de datos junto a comunidades para cerrar vacíos críticos de datos. También se requiere diversidad en los equipos que diseñan y despliegan tecnologías, para mitigar sesgos invisibles en equipos homogéneos. Finalmente, hace falta una mirada de sostenibilidad que contemple impactos ambientales y materiales (energía, agua, hardware, cadenas extractivas), condicionando despliegues a criterios de eficiencia y adecuación al propósito.

En este sentido, la sociedad civil ha sido clave para instalar estas dimensiones en la conversación pública y para generar herramientas de evaluación, desde guías de métricas inclusivas hasta protocolos de consulta y evaluación de impacto diferenciado.

Las consecuencias de perpetuar un modelo de transparencia mínima y gobernanza débil son demasiado altas. Corremos el riesgo de ampliar la discriminación algorítmica, consolidar la vigilancia desproporcionada, especialmente sobre defensores de derechos humanos, y normalizar decisiones opacas e inapelables. Los bloqueos tecnológicos y contractuales pueden inmovilizar a las administraciones, encarecer el ciclo de vida de los sistemas y dificultar las correcciones. Y, en el plano democrático, la confianza pública puede erosionarse.

Para evitar este escenario, necesitamos una gobernanza de la IA que asegure sistemas respaldados por trazabilidad y mejora continua; capacidad institucional para auditar, corregir y retirar sistemas; una innovación ética desde el diseño, con datos y modelos adecuados a problemas locales; y cohesión democrática fortalecida por procesos abiertos y rendición de cuentas.

Conclusión

El índice global de IA responsable revela un escenario que América Latina no puede ignorar: la región innova, pero las regulaciones aún están fragmentadas e insuficientes. En esta situación, la sociedad civil y la academia están sosteniendo y empujando la agenda que ancla la IA al marco de derechos humanos. De esta manera, marcan un camino hacia dónde debemos avanzar: dispositivos de gobernanza de los datos y la IA que sean multiactor y prioricen un

enfoque de diversidad e inclusión de las distintas comunidades vulnerables ante la IA, asegurando que estos sistemas contribuyan al bien común y no reproduzcan desigualdades históricas.

Esta gobernanza democrática de la IA requiere dispositivos que vuelvan exigibles los principios: transparencias que permitan comprender y monitorear; auditorías con acceso real; rutas efectivas de reparación; compras públicas que condicionen la provisión; participación significativa que modifique decisiones; y métricas de equidad que capten riesgos y efectos. En cada una de estas discusiones, la sociedad civil latinoamericana ha demostrado capacidad para diagnosticar, diseñar y acompañar la implementación y, cuando es necesario, activar contrapesos.

Ese es, en América Latina, el núcleo de una agenda de gobernanza que no solo gestiona riesgos, sino que habilita futuros democráticos más justos.

Referencias

- Artículo 19, Thomson Reuters Foundation. (2024). Hacia una transparencia algorítmica en el sector público de América Latina. Su adquisición, implementación y desarrollo.
- Global Index on Responsible AI. (2024). Global Index on Responsible AI 2024 (Edición corregida).
- GPAI/OCDE. (2025). Transparencia algorítmica en el sector público: estado del arte de instrumentos.
- UNESCO. (2021). Recomendación sobre la Ética de la Inteligencia Artificial.

En torno a las decisiones autónomas de la Inteligencia Artificial: libertad, juicio y responsabilidad

Autor: Mario A. Sandoval M. ITESM-CCM/FFYL UNAM

"El mayor peligro no es la técnica misma, sino que el hombre se vea atrapado en la ilusión de dominarla." Martin Heidegger, La pregunta por la técnica

Sin duda uno de los lemas que selló el pensamiento de la filosofía moderna fue la frase que Kant acuñó en su texto "¿Qué es ilustración?" ¡Sapere Aude! ¡Ten el valor de servirte de tu propio entendimiento! Bajo el auspicio de este lema, el pensamiento moderno dio inicio a una época caracterizada por la entronización de la razón y una promesa que implicaba el dominio y control del mundo en favor del ser humano. La razón aparece como garante del inicio de un camino hacia el cosmopolitismo, que en palabras de nuestro autor, es también el destino legal de la humanidad.

¿Qué caracteriza la salida de esta minoría de edad que Kant enuncia? El uso de la razón por cuenta propia. Sin temor a equivocarnos, las pretensiones planteadas por el autor alemán hoy podrían cuestionarse bajo un juicio diferente al que siempre se le ha sometido. La conmoción histórica que causó la instrumentalización de la razón llevó a autores como Horkheimer y Adorno a llamar a la ilustración la triunfal calamidad (Horkheimer, p. 59). Dichos pensadores nos invitan a reflexionar en torno a las implicaciones que la triada razón, poder y dominio trajeron a mediados del siglo XX. Los autores de la llamada Escuela de Frankfurt lo enuncian de la siguiente manera: "Sin consideración para consigo misma, la ilustración ha consumido el último resto de su propia autoconciencia" (Ibid., p. 60). A poco más de ochenta años de publicada *La Dialéctica de la Ilustración*, sus reflexiones cimbran nuevamente el estado del mundo actual que ahora cede parte de los procesos deliberativos en manos de la inteligencia artificial.

La capacidad de elección es una condición humana, un existenciario que nos constituye. Es una cualidad que no puede ser cancelada porque hacerlo sería un atentado contra la dignidad. Sin embargo,

poco a poco y de manera voluntaria hemos empezado a dejar decisiones a las diversas inteligencias artificiales. Sin caer en una demonización de ellas, pretendemos abordar desde un ámbito filosófico las implicaciones éticas que tiene esta situación coyuntural en nuestro mundo contemporáneo. Sabemos que las reflexiones vertidas aquí son tan sólo líneas de pensamiento en torno a una situación que apremia a ser pensada desde diversas aristas. Este texto contribuye, así, a ese diálogo colectivo.

¿Es posible una ética de la inteligencia artificial que preserve la pluralidad, el juicio y la libertad humana en la toma de decisiones?

De acuerdo con varios filósofos de la posguerra, uno de los elementos que condujo a la barbarie más conocida del siglo XX es justo el factor de la elección o, dicho de otra manera, el no hacerse responsable de las propias elecciones. Sartre, en *El existencialismo es un humanismo*, menciona: "Es lo que expresaré diciendo que el hombre está condenado a ser libre. Condenado, porque no se ha creado a sí mismo y sin embargo, por otro lado, libre, porque una vez arrojado al mundo es responsable de todo lo que hace" (Sartre, 1999, p. 43). Es la libertad de poder elegir y hacer efectiva esa capacidad lo que hace auténtica la vida humana. Negarla o cancelarla sume a la vida en una inautenticidad. Nuestro autor llama a esto mala fe. Cancelar la posibilidad de pensar es actuar de mala fe y con ello la cancelación de libertad.

Arendt, en *La condición humana*, hace énfasis en la importancia del juicio crítico para la acción colectiva. El ser humano es un ser político que construye y determina la dignidad a partir de la acción pública. Dejar de lado el ejercicio deliberativo, el ejercicio de elección entre pares, es un atentado contra esa dignidad. No decidir nos deja frente a una renuncia al pensar y al juzgar en el día a día. En otra de sus grandes obras, *La banalidad del mal*, alerta sobre la pérdida de la capacidad de elección. Eichmann es un personaje que no tiene un diálogo interior consigo mismo y tampoco debate públicamente sobre las consecuencias indeseadas de sus actos. Las consecuencias son conocidas por todos.

En el texto *No-cosas* (Han, 2022), el autor describe nuestra época como la era de transición de los objetos a la información digital. Ésta presenta un carácter inmaterial, inasible. Sin embargo, esa inmaterialidad es suplantada por una cascada de datos digitales cuya única mediación en su flujo son los algoritmos. "Es la información, no las cosas, la que determina el mundo en que vivimos. Ya no habitamos la tierra y el cielo, sino Google Earth y la nube" (Han, 2022, p. 13). Este fenómeno inaugura una era en la

que la mediación tecnológica redefine los límites de la acción humana.

El agobio de interfaces y datos no es tema baladí: la información determina grandes parcelas del mundo en el que vivimos. Educación, economía y relaciones sociales están determinadas por el flujo incesante de información digital a la que accedemos. Nuestro ser cotidiano se mueve entre datos que nutren la infoesfera, ese espacio donde nos comunicamos e intercambiamos información con los artefactos inteligentes. Según el GIRAI 2024, "la inteligencia artificial responsable debe diseñarse, desarrollarse y gobernarse de manera que respete y proteja los derechos humanos y defienda los principios de la ética en cada etapa del ciclo de vida de la IA" (Global Index on Responsible AI, 2024, p. 5). Este recordatorio marca el punto de partida para entender que toda mediación tecnológica debe estar subordinada a fines humanos y no al revés.

Cuando el uso de la inteligencia artificial alcanza ámbitos como la educación o las finanzas sin un control responsable, su empleo sufre condiciones deformativas. En esa conformación caracterizada por lo recién enunciado, la inteligencia artificial debe empezar a tomar decisiones. Se generan conflictos de corte ético, entre los que están el sacrificio de la autonomía individual, las tensiones en términos de equidad y los límites de acceso a la información. En este punto, el GIRAI 2024 enfatiza la necesidad de "promover sistemas de IA inclusivos, transparentes y auditables que permitan identificar y mitigar los sesgos que puedan generar daño social" (Global Index on Responsible AI, 2024, p. 8). No hacerlo puede ser un atentado a la justicia social.

Y es que en cierta medida podría parecer que la IA alivia esa tensión por tener que elegir, pero sin duda y siguiendo nuevamente a Sartre, la condena a elegir es un existenciario que caracteriza a los seres humanos. Dejar que la máquina decida no nos exime de la responsabilidad moral por sus resultados, sino que la amplía al ámbito de su diseño, control y supervisión.

Los resultados algorítmicos no son sino una constante repetición de datos. Están sesgados a lo que la mayoría piensa, pero eso no implica que tengan elementos de justicia. Algunas veces replican las desigualdades y continúan con la segregación a grupos históricamente excluidos. En relación con esto, el GIRAI 2024 advierte que "la inteligencia artificial debe desarrollarse bajo principios de justicia y no discriminación, garantizando que las tecnologías no reproduzcan inequidades estructurales" (Global Index on Responsible AI, 2024, p. 10). De este

modo, el documento ofrece un marco ético que exige mantener la vigilancia sobre los efectos sociales de la IA, no cancelando su uso pero si restringiéndolo a condiciones de equidad y justicia.

Confiar ciegamente en la tecnología generativa conduce a lo que Heidegger llama encantamiento. Producto de la técnica moderna, el ser humano es sometido a un poder que lo ata, lo retiene y lo configura. El encantamiento es producto "del desenfrenado dominio de la maquinación [...]. El hechizo de la técnica y sus progresos que permanentemente se aventajan entre sí, son sólo un signo de este encantamiento, que en consecuencia impele todo a cálculo, utilización, cultivo, manejabilidad y regulación". De ahí la fascinación que la tecnología genera mediante los conceptos de una realidad que engrandece conceptos como eficiencia, el dominio y el poder de transformación inmediata. Pero lo cierto es que encierra al ser humano en un sistema técnico del cual se vuelve dependiente. El GIRAI 2024 alerta que "la automatización sin supervisión humana erosiona la capacidad crítica y reduce el espacio para el juicio ético" (Global Index on Responsible AI, 2024, p. 12), señalando así la necesidad de preservar el papel deliberativo de las personas en los sistemas tecnológicos.

¿Hasta qué punto es deseable plantear la cuestión y la deliberación en torno al papel de la libertad y el juicio propio? De entrada consideramos que el uso ilimitado de la IA puede generar un desapego del ser humano de la realidad concreta. Pensemos en las implicaciones de estas prácticas: una disolución de la capacidad crítica y la formación de una mera perspectiva ética a la cual podríamos llamar algorítmica. La facticidad reúne las características concretas que constituyen la identidad y autonomía de la vida humana. El uso desmedido de las IA genera un alejamiento de dicha facticidad. Un ser humano alejado de la vida, de la realidad, cancela la posibilidad de cualquier tipo de juicio. En este sentido, el GIRAI 2024 propone que "las decisiones automatizadas deben ser siempre interpretables y revisables por humanos, garantizando la transparencia y la rendición de cuentas" (Global Index on Responsible AI, 2024, p. 14). Esta afirmación dialoga directamente con la advertencia de Han sobre la pérdida de facticidad, mostrando que la ética de la IA solo es posible si se preserva la mediación humana en toda decisión.

Jonas propone, con su principio de responsabilidad, que el ser humano debe prever las consecuencias de sus actos sobre la vida futura. Si somos consecuentes con lo mencionado hasta ahora, una "ética algorítmica", además de

reproducir y perpetuar sesgos al tomar decisiones basadas meramente en la repetición, diluye el papel de la responsabilidad al mostrar decisiones basadas en la abstracción de la realidad. El algoritmo borra el rostro del otro y lo sustituye por perfiles y datos que amplían la vulnerabilidad de grupos históricamente sojuzgados. El otro se vuelve un halo flotante, un dato o número encriptado en un código binario que carece de reconocimiento. Frente a ello, el GIRAI 2024 enfatiza la importancia de "preservar la agencia humana, la diversidad cultural y la inclusión de voces múltiples en la gobernanza de la inteligencia artificial" (Global Index on Responsible AI, 2024, p. 16).

De este modo, nuestra época cargada de la necesidad de resultados ha delegado muchos elementos decisionales a las IA. Sin percatarnos, los fundamentos de la libertad y de la responsabilidad son puestos en tensión. La libertad no puede restringirse a las condiciones de elección de plataformas o a la creación de mejores prompts que den lugar a respuestas más precisas siempre en relación con los parámetros dictados por una realidad meramente eficiente e instrumental. La responsabilidad hoy en día no puede limitarse a responder por los actos que hacemos, sino también por las omisiones morales que causamos y por un pensar que no es desarrollado y alimentado. En palabras del GIRAI 2024, "el desarrollo responsable de la IA requiere un compromiso ético continuo que integre valores humanos, sostenibilidad y justicia en todas las etapas del proceso" (Global Index on Responsible AI, 2024, p. 18).

El uso de la IA debe implicar siempre pensar en términos de justicia, pues el riesgo de invisibilizar al otro es siempre manifiesto. Se debe pensar que la esencia humana es la acción fáctica: ser en el mundo, ser con los otros. La IA sin mediación, sin condiciones críticas en relación con su uso, aleja al ser humano del mundo poniendo en tela de juicio nuestra propia libertad: libertad a ser, libertad de pensar, libertad de elegir.

Insistimos: no se trata de demonizar la inteligencia artificial, sino de hacer patente que su uso requiere marcos éticos, políticos y normativos firmes. La tecnología, en tanto herramienta potente, amplifica decisiones y efectos; no los neutraliza. Como recuerda el GIRAI 2024, "es a través del trabajo conjunto desde pericias y puntos de vista diversos que podemos impulsar decisiones verdaderamente responsables" (Global Index on Responsible AI, 2024, p. 20). Ese llamado a la deliberación plural es, precisamente, el punto de apoyo para construir una ética que no se reduzca a protocolos técnicos o líneas de ornato en códigos de escuelas y oficinas. Lo que aquí se juega es el juicio humano, la responsabilidad y el cuidado del futuro.

Reiteramos que ceder las decisiones a cajas negras algorítmicas, no elimina la exigencia ética. Quien diseña, implementa o legítima sistemas de decisión sigue siendo responsable de sus efectos y omisiones. Recordemos lo que Kant planteaba en relación con la buena intención: no basta con ella, hacen falta espacios públicos donde el juicio se ejerza y sea exigible.

Disolver lo concreto, la facticidad de lo encarnado, de lo situado, tiene consecuencias éticas directas: sin anclajes factuales, la deliberación se empobrece y la responsabilidad se vuelve abstracta. El pensar contemporáneo debe salir de su instrumentalidad y mirar más allá del presente inmediato para proteger el pensamiento humano frente a riesgos tecnológicos enormes. La delegación algorítmica sin previsión ni obligaciones hacia el porvenir vulnera ese principio: decidir hoy sin pensar en mañana es, en términos jonians, éticamente insostenible.

Levinas insistió sobre el riesgo de la reducción del rostro del otro a meros datos: la ética surge del reconocimiento y se realiza en la responsabilidad incondicional hacia él. Si los sistemas solo leen perfiles, no rostros, corremos el riesgo de normalizar una ética ciega al sufrimiento concreto. Se deben reconocer las vulnerabilidades sociales no como una repetición numérica de datos, sino como factualidades que atañen a todos.

La tarea es grande, nos lleva a no sólo mantenernos en el ámbito reflexivo que la filosofía y la ética plantean. Sino llevar a hechos varias implicaciones normativas y operativas que corresponde pensar a otros especialistas y encargados de la formación de jóvenes.

Compartimos los esfuerzos que son enunciados en el GIRAI y lo reconocemos como un esfuerzo importante para pensar lo aún no pensado en relación con estas tecnologías. Finalmente, el reto ético de nuestra era no trata de frenar la innovación, sino de inscribirla en procedimientos de juicio y responsabilidad que preserven la dignidad humana, la pluralidad de voces y la continuidad del pensamiento. Recuperar la facticidad frente al algoritmo implica, en definitiva, recuperar la exigencia de pensar y de responder: condiciones ineludibles para que la tecnología sea aliada de lo humano y no su sustituta.

Referencias

- Arendt, H. (2003). Responsabilidad y juicio (M. A. Bermejo, Trad.). Paidós.
- Beauvoir, S. de. (2006). La ética de la ambigüedad (M. F. García, Trad.). Edhasa.

- Global Index on Responsible AI. (2024). GIRAI 2024: Informe sobre inteligencia artificial responsable [Informe en línea]. Recuperado de https://www.global-index.ai/
- Han, B.-C. (2022). No-cosas. Quiebras en el mundo de hoy (A. M. Sánchez, Trad.). Taurus.
- Heidegger, M. (2003). Aportes a la filosofía. Acerca del evento. (Dina V. Picotti, Trad). Biblos
- Jonas, H. (1995). El principio de responsabilidad: Ensayo de una ética para la civilización tecnológica (J. M. Navarro Cordón, Trad.). Herder.
- Sartre, J.-P. (1999). El existencialismo es un humanismo (V. Molina, Trad.). Alianza Editorial.

JÓVENES + IA: PERSPECTIVAS DE UNA GENERACIÓN
JÓVENES + IA: PERSPECTIVAS DE UNA GENERACIÓN

JÓVENES + IA

LOS DILEMAS ÉTICOS DE LA IA

A continuación, presentamos los ensayos escritos por diez alumnos y alumnas que participaron en nuestra convocatoria abierta sobre ética, datos e inteligencia artificial. Esta convocatoria invitó a estudiantes universitarios del Tecnológico de Monterrey a reflexionar críticamente sobre los desafíos éticos vinculados con el uso de los datos y el desarrollo de sistemas de IA en nuestras sociedades.

De todas las propuestas recibidas, se seleccionaron diez trabajos que destacaron por su rigor analítico, claridad argumentativa y capacidad para conectar los dilemas tecnológicos actuales con debates más amplios sobre derechos humanos, transparencia, equidad y justicia social.

Estos ensayos, elaborados basándose en el *Global Index on Responsible AI 2024*, muestran la importancia de incorporar la voz y la perspectiva de las nuevas generaciones en la discusión sobre el presente y el futuro de la IA. Cada texto constituye una contribución valiosa para comprender cómo los sistemas automatizados pueden impactar positiva o negativamente nuestras vidas y nuestras instituciones.

El Riesgo de la Neutralidad Algorítmica: Ética y Poder en las Decisiones Automatizadas

Autores: Ana Ana Rello de Obeso y Diego Alberto Carrillo Castro

La delegación de decisiones críticas a sistemas automatizados es un fenómeno que ha transformado nuestra habilidad de toma de decisiones que vino en conjunto con la Inteligencia Artificial. Por ejemplo en programación ya se tiene el *método greedy* que encuentra una solución globalmente óptima a un problema a base de hacer elecciones localmente óptimas (Russell, Norvig, 2021). En palabras más sencillas, el algoritmo siempre hace lo que "parece" mejor en cada momento, jamás reconsidera y acaba llegando a la mejor solución posible que el programador haya denominado como tal. En la vida cotidiana, esto puede hacerse con circunstancias donde se debe de asignar los recursos médicos a un hospital, conceder un préstamo, realizar una contratación, entre muchas otras cuestiones que antes eran exclusivamente humanas pero ahora los diversos algoritmos asumen más responsabilidades.

Esta lógica de optimización inmediata también se encuentra en muchos sistemas actuales de inteligencia artificial aplicados a la toma de decisiones automatizadas. Los llamados sistemas de soporte automatizado y las herramientas del análisis predictivo operan bajo esta misma premisa: procesan grandes volúmenes de datos y producen una salida que, sin contemplar el panorama general, busca la mejor decisión posible dentro de un marco definido por los programadores. Estos resultados pueden ir desde recomendaciones triviales ya sea como "este restaurante coincide con tus preferencias" hasta decisiones críticas como negar un préstamo, rechazar una solicitud de empleo, asignar un órgano a otro paciente o incluso identificar un blanco militar. Uno de los usos más polémicos de este tipo de predicción se encuentra en la llamada "policía predictiva", donde, en lugar de esperar que ocurra un delito, se intenta anticiparlo con base en patrones estadísticos (Müller, 2023). Críticos como Ferguson (2017) advierten que esta práctica puede erosionar libertades civiles, al trasladar el poder de decisión desde los ciudadanos hacia modelos algorítmicos que no rinden cuentas.

Además, estas tecnologías tienden a reproducir sesgos históricos, como lo evidenció el sistema COMPAS, que mostró una mayor tasa de falsos positivos en personas negras. Esta situación revela una preocupación ética clave: aunque se han desarrollado esfuerzos técnicos para mitigar el sesgo en la inteligencia artificial, estos aún son limitados, pues requieren conceptos como "justicia" o "raza" definidos en términos matemáticos, lo cual resulta problemático. Así, nos enfrentamos no solo a una cuestión de eficiencia o precisión técnica, sino a una posible transferencia injustificada de autoridad moral a sistemas que no poseen agencia, empatía ni capacidad de deliberación.

En *The Oxford Handbook of Ethics of AI* (2020), se advierte que muchas decisiones automatizadas, especialmente en contextos críticos como la justicia, la salud o la asignación de recursos sociales, carecen de mecanismos claros de rendición de cuentas. La delegación de autoridad a sistemas algorítmicos opacos pone en riesgo principios fundamentales del orden democrático, como la transparencia, la equidad y el derecho a apelar decisiones que afectan directamente la vida de las personas. Además, se resalta cómo estos sistemas pueden reproducir y amplificar sesgos estructurales presentes en los datos de entrenamiento, perpetuando desigualdades históricas bajo una apariencia de neutralidad técnica. Este fenómeno no es accidental, sino resultado de un diseño que muchas veces omite la reflexión ética y la participación de grupos diversos en su desarrollo. Además de que agente en el contexto de la Inteligencia Artificial es aquella entidad que recibe percepciones del entorno y realiza acciones, las cuales se realizan de manera mecánica en función de una secuencia de percepciones; sus objetivos son externos: no definen intenciones propias ni comprenden sus actos y estos no tienen conciencia ni iniciativa propia. (Dubber, Pasquale, & Das, 2020)

Por su parte, Kate Crawford, en *The Atlas of AI* (2021), plantea una crítica aún más radical al afirmar que la inteligencia artificial no es simplemente una herramienta neutral, sino una tecnología profundamente imbricada en relaciones históricas de poder, dominación y extracción de valor. Desde la recolección de minerales necesarios para la infraestructura de hardware hasta la explotación de datos humanos como insumo de entrenamiento, la IA forma parte de un sistema extractivista global que beneficia a unos pocos y profundiza las desigualdades económicas, raciales y de género. En este marco, la IA no solo plantea desafíos técnicos, sino también éticas de justicia estructural y responsabilidad colectiva. Esta tendencia obliga a replantear críticamente los

discursos de eficiencia y objetividad que suelen acompañar a estas tecnologías, pues encubren procesos que pueden llevar a una pérdida significativa de agencia humana, tanto en el plano individual como social. En suma, estas obras coinciden en señalar que la creciente automatización de decisiones plantea interrogantes urgentes sobre la responsabilidad moral de los agentes humanos, la necesidad de transparencia algorítmica, y la posibilidad de que se institucionalice una forma de discriminación automatizada bajo el ropaje de la innovación tecnológica.

Es por ello que se cuestiona esta situación como problema moral. ¿Es éticamente justificable delegar decisiones críticas a sistemas automatizados, incluso cuando estas pueden tener un impacto irreversible sobre la vida de las personas?

Por un lado, se podría responder el dilema con la postura de que sí, es justificable si el sistema ha demostrado mayor precisión y objetividad que los humanos, ya que los sistemas automatizados pueden analizar grandes cantidades de datos sin sesgos emocionales o fatiga. Por otra parte, también se debe de considerar el tipo de decisión y del contexto, ya que algunas decisiones pueden ser delegadas (como asignación de recursos médicos) pero no aquellas donde existen dilemas éticos complejos.

De la mano, nuestro dilema ético sería la siguiente pregunta: ¿Deberíamos seguir utilizando sistemas automatizados para tomar decisiones críticas si ofrecen eficiencia y coherencia, a pesar del riesgo de errores y sesgos, o deberíamos limitar su uso en favor de la intervención humana, aunque esto implique procesos más lentos o subjetivos?

Para un análisis más completo, es necesario comprender la *accountability* en el contexto de la inteligencia artificial, al igual que definir conceptos clave como *ownership* y *accountability*, según lo propuesto por diversos autores como Dubber, Pasquale y Das (2020). En este contexto, *ownership* se refiere a la propiedad cognitiva, es decir, la capacidad de un agente (ya sea un ser humano o, hipotéticamente, una inteligencia artificial) para apropiarse de un plan de vida, un conjunto de creencias, valores o deseos, de manera que estos elementos formen parte integral de su identidad y autonomía. En el caso de los sistemas de inteligencia artificial, aunque no poseen conciencia ni autonomía plena, los agentes humanos responsables de su diseño y operación deben asumir este tipo de ownership respecto a las metas y valores incorporados en dichos sistemas.

Por otro lado, accountability en Inteligencia Artificial implica la obligación de los agentes con *ownership* (los

desarrolladores, operadores y organizaciones) de rendir cuentas por las acciones y consecuencias generadas por los sistemas que crean y despliegan. Esto requiere establecer líneas claras de responsabilidad y mecanismos que permitan evaluar cuándo y cómo un agente debe ser considerado responsable por violaciones a normas o valores sociales. La *accountability* actúa como un mecanismo para abordar la complejidad de definir valores normativos abstractos (como justicia o equidad) y, a la vez, para supervisar y corregir comportamientos problemáticos en sistemas automatizados. La transparencia, entendida como el acceso a información sobre el funcionamiento del sistema, es un recurso instrumental para la accountability, pero no puede reemplazarla.

Es por ello que la gobernanza de la Inteligencia Artificial debe fundamentarse en estos conceptos para garantizar que los agentes humanos que crean y controlan estas tecnologías puedan ser responsabilizados efectivamente. Esto implica no solo la aplicación de herramientas técnicas (como registros criptográficos y verificaciones de software), sino también la implementación de procesos organizativos rigurosos, que incluyan documentación detallada, revisiones independientes y evaluación continua. De este modo, la *accountability* se consolida como un mecanismo esencial para alinear el desarrollo de la inteligencia artificial con normas democráticas y valores humanos, asegurando que, aunque las máquinas carezcan de autonomía moral, los agentes humanos mantengan la propiedad cognitiva y la responsabilidad ética de sus creaciones.

La ética de Aristóteles, también conocida como teoría de la virtud, sostiene que el fin último del ser humano es alcanzar la eudaimonía, entendida no como placer inmediato, sino como una vida plena, en la que el individuo realiza su potencial humano a través del cultivo de virtudes morales. Estas virtudes no son innatas, sino hábitos de elección adquiridos mediante la práctica y la corrección, que permiten actuar correctamente en la vida cotidiana. (Sánches Hernández, 2002)

Las virtudes, según Aristóteles, se sitúan en un punto medio entre dos extremos viciosos: uno por exceso y otro por defecto. Por ejemplo, la virtud de la valentía se encuentra entre la temeridad y la cobardía. Este principio del justo medio se aplica también al uso de la inteligencia artificial: debemos evitar tanto el entusiasmo desmedido y ciego por la automatización como el rechazo total a su utilidad. (Sánches Hernández, 2002)

Asimismo, Aristóteles subraya que la

virtud es siempre contextual: no basta con actuar con buenas intenciones; se debe actuar como, cuando, donde y con quien es preciso hacerlo, lo cual exige juicio moral y sensibilidad a las circunstancias concretas. Y, algo clave en este análisis: la evaluación moral de un acto requiere atribuir responsabilidad al agente, lo que implica que debe haber voluntariedad y deliberación ética en las decisiones.

Una postura que se puede tener ante este dilema ético es aquella donde la automatización es vista como expresión de virtud prudente y justa. Para la justificación de esta, desde esta perspectiva aristotélica, el uso de sistemas automatizados en decisiones críticas puede ser ético si se realiza con virtudes como la prudencia, la responsabilidad y la justicia. La prudencia (*phronesis*) permite deliberar sobre lo que es bueno en cada situación, integrando el conocimiento técnico con la sabiduría moral. La justicia (la cual se entiende como la virtud que contiene a todas las demás) implica diseñar sistemas que no reproduzcan desigualdades y respeten la dignidad de todas las personas.

Una manera de ilustrar esta postura es la siguiente: en un caso hipotético se tiene a un grupo de personas interdisciplinarias que diseña un sistema de IA para asignar recursos médicos en situaciones de emergencia. Este grupo revisa constantemente el funcionamiento del sistema, permite la intervención humana en decisiones críticas y capacita al personal para actuar con sensibilidad ética. Además asume la responsabilidad ética del proceso y técnica por las decisiones que el sistema toma, asegurando que siempre exista un agente humano que pueda responder ante las consecuencias. Su acción refleja virtudes como la templanza (no depender ciegamente del algoritmo), la justicia (priorizar a quienes más lo necesitan) y la fortaleza (afrontar los dilemas morales que surgen en situaciones extremas), lo cual encarna un enfoque basado en las disposiciones virtuosas de los agentes involucrados.

Este ejemplo no se aleja de la realidad, ya que el sistema eTriage, desarrollado por la Universidad de Sheffield junto con la Fundación NHS de los Hospitales Universitarios de Sheffield, muestra cómo la inteligencia artificial puede ser implementada con responsabilidad ética en el ámbito médico. Utilizando un algoritmo de aprendizaje automático entrenado en más de 600.000 registros, eTriage permite predecir la gravedad de los pacientes en función de síntomas, signos vitales e historial médico, ayudando a priorizar los casos más críticos y a asignar recursos de forma justa y eficaz. Este tipo de innovación tecnológica no solo optimiza la atención en emergencias, sino que también

demuestra cómo la virtud en la acción de los diseñadores y usuarios de la IA puede dar lugar a sistemas más humanos y confiables (Clemente et al., 2024).

Sin embargo, la otra cara de la moneda es la postura donde la automatización es vista sin virtud como falta de responsabilidad y sensibilidad moral. Esto puede llegar a suceder al momento de delegar decisiones críticas a sistemas automatizados sin reflexión ética puede reflejar vicios morales, como la indiferencia, la imprudencia o la cobardía ética. En lugar de ejercer juicio, se prefiere dejar la responsabilidad a un sistema, lo cual contradice la idea aristotélica de que la virtud requiere voluntad y responsabilidad moral.

Una ejemplificación de este caso sería cuando una empresa adopta un algoritmo para decidir qué pacientes recibirán tratamientos prioritarios, sin evaluar los sesgos en los datos ni considerar apelaciones humanas. Esto evidencia una falta de empatía, de prudencia y de justicia, ya que el sistema reproduce desigualdades históricas y deja a personas vulnerables sin acceso a atención. El agente moral (la empresa o sus responsables) no actúa con virtud, sino que se escuda en la "objetividad" de la tecnología para evitar cuestionamientos éticos. Si se usara el eTriage sin supervisión humana, sería un caso donde la automatización faltaría de virtud, responsabilidad y sensibilidad moral.

Este debate sobre las posturas éticas no solo es relevante en el plano filosófico, sino que cobra especial importancia cuando se observa desde una perspectiva de derechos humanos. Como lo indica el *Global Index Report on Responsible AI (*2024), en la mayoría de los países aún no existen condiciones adecuadas para avanzar hacia sistemas de inteligencia artificial que sean inclusivos y equitativos. Los temas relacionados con los derechos de grupos marginados o históricamente desatendidos están entre los peor evaluados, lo que evidencia una falta de prioridad gubernamental en cuanto a la equidad y la inclusión en la IA. Por ejemplo, la protección de los derechos de la infancia solo es considerada adecuadamente en los países con mejor desempeño, mientras que la discriminación y los sesgos no figuran entre las áreas temáticas más fuertes, incluso en esos mismos países. Esto confirma que, sin un marco ético robusto que guíe el diseño, la implementación y el uso de sistemas automatizados, se corre el riesgo de consolidar desigualdades existentes. La ética de la responsabilidad y la ética del cuidado no son, por tanto, una opción accesoria, sino una condición indispensable para garantizar que la tecnología respete y promueva la dignidad humana.

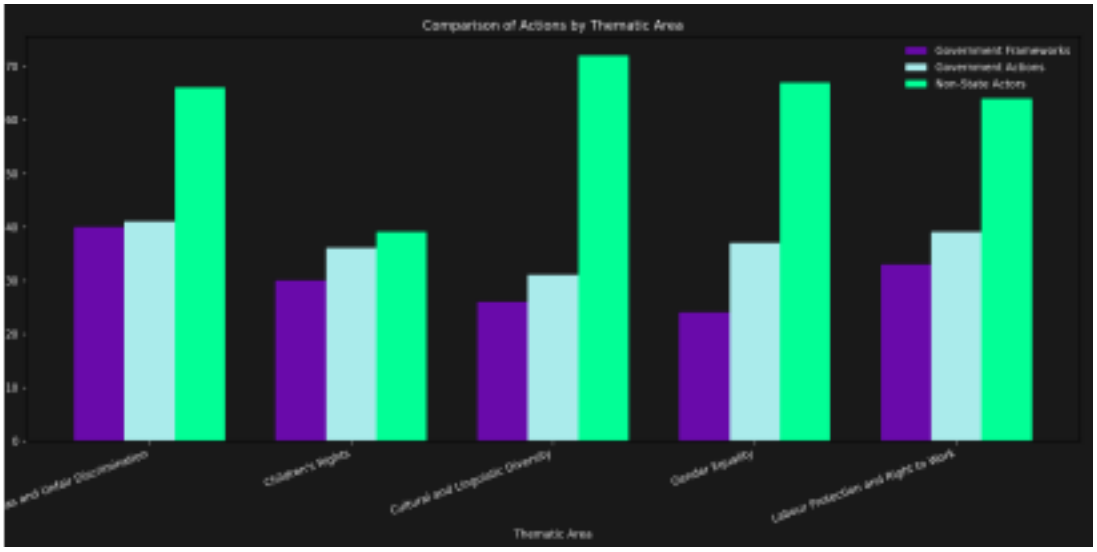


Figura 1. *Gráfico de número de países con evidencia en las áreas temáticas relacionadas con la IA y la inclusión por pilar*, 2024.[1]

El gráfico muestra el número de países que presentan evidencia en diferentes áreas temáticas vinculadas a la inteligencia artificial y la inclusión. Este destaca que en todas las temáticas (como la discriminación, los derechos de la infancia, la diversidad cultural y lingüística, la igualdad de género y la protección laboral) los actores no estatales son quienes más evidencias aportan. Particularmente relevante es el caso de la diversidad cultural y lingüística, donde 72 países tienen participación activa de estos actores, muy por encima de los marcos y acciones estatales. Esto sugiere que la sociedad civil y el sector privado están desempeñando un papel crucial en la promoción de la inclusión en la IA, frente a una respuesta gubernamental aún limitada o desigual.

Este panorama puede ser interpretado desde la ética de la virtud basada en el agente, la cual sostiene que la normatividad moral no se deriva de estados externos como la utilidad o el bienestar, sino de las disposiciones motivacionales de los agentes. A la luz de esta perspectiva, la limitada participación gubernamental en la promoción de una IA inclusiva no solo revela una falla política o institucional, sino una insuficiencia en las motivaciones éticas de quienes están en posiciones de poder.

Según Slote (1993), una acción es moralmente buena si responde a buenas motivaciones, como la empatía o la compasión. Así, la falta de acciones efectivas en temas como la discriminación, los derechos de la infancia o la diversidad cultural evidencia que los responsables de políticas públicas carecen de disposiciones virtuosas suficientes para priorizar el bienestar de los grupos más vulnerables. Por el contrario, el papel activo que juegan actores no estatales en la promoción de la equidad en la IA puede entenderse como una manifestación de motivaciones virtuosas ejemplares. Desde el enfoque exemplarista de Zagzebski, identificamos como moralmente valiosas aquellas acciones que emulan las disposiciones de agentes virtuosos, que actúan con responsabilidad, sensibilidad y compromiso ético incluso cuando no están obligados a hacerlo. Estos actores (sociedad civil, académicos, sector privado ético) se convierten así en referentes morales (*exemplars*), no porque sigan reglas externas, sino porque su actuar refleja una disposición coherente con lo que entendemos como virtud. Su compromiso con la inclusión tecnológica no solo marca una diferencia práctica, sino que sirve como guía para juzgar la insuficiencia moral de las instituciones públicas. De este modo, el agente virtuoso no solo actúa correctamente, sino que encarna un modelo que revela, por contraste, el carácter deficiente de quienes omiten actuar.

Como advierte Hannah Arendt (2003), la estructura burocrática de las instituciones modernas contribuye a diluir la responsabilidad moral, al fragmentar las decisiones en una red de procedimientos impersonales donde ningún agente asume plena responsabilidad por las consecuencias. De este modo, no es solo que se esté haciendo poco, sino que lo que se hace parece impulsado por motivos inadecuados o insuficientemente buenos y, al mismo tiempo, legitimado por dinámicas institucionales que despersonalizan la acción moral, lo cual, según este marco teórico, vuelve aún más cuestionable la inacción institucional.

Ambas posturas sobre la automatización reflejan tensiones éticas profundas y relevantes en el debate actual sobre la inteligencia artificial. Por un lado, la visión optimista de la automatización como una expresión de virtud se fundamenta en la idea de que las tecnologías pueden ser diseñadas, implementadas y supervisadas por agentes virtuosos, cuyas motivaciones están guiadas por la prudencia, la justicia y el compromiso con el bienestar común. En este marco, los sistemas de IA no sustituyen el juicio humano, sino que lo amplifican; son herramientas que ayudan a tomar

decisiones más informadas, rápidas y eficaces, como ocurre en el caso del sistema eTriage. La virtud, en este caso, se manifiesta no solo en la eficiencia tecnológica, sino en la disposición del agente a mantenerse alerta, corregir fallas y priorizar éticamente los recursos limitados.

En contraste, la postura crítica advierte que la automatización sin virtud puede convertirse en una forma de deshumanización: cuando los agentes se desentienden de su responsabilidad moral o se apoyan ciegamente en decisiones automatizadas, se pierde la sensibilidad ética que requiere cada situación concreta. Desde una perspectiva de ética de la virtud basada en agentes, como la de Zagzebski o Slote, esto implica que los sistemas técnicos no son problemáticos por sí mismos, sino por la falta de cualidades morales en quienes los diseñan y utilizan. Un agente indiferente, guiado por la eficiencia sin empatía ni juicio, no actúa virtuosamente, aunque su acción sea técnicamente correcta. Esta ausencia de virtudes como la compasión, la templanza o la responsabilidad puede consolidar injusticias preexistentes, profundizar sesgos o invisibilizar sufrimientos.

Ante este dilema, una solución razonada desde la ética de la virtud debe buscar el equilibrio: se trata de formar e identificar agentes capaces de actuar virtuosamente dentro de sistemas tecnológicos complejos. Esto implica desarrollar marcos institucionales que no solo regulen la IA desde parámetros técnicos o legales, sino que fomenten la formación de carácter ético en los diseñadores, usuarios y tomadores de decisión. Una ética de la virtud aplicada a la IA no solo pregunta "¿qué hace esta tecnología?", sino "¿quién la hace, por qué, y con qué disposición moral?" de acuerdo con la teoría presentada por Slote (1993). La solución no está en rechazar la automatización ni en abrazarla ciegamente, sino en cultivar agentes ejemplares cuya motivación no sea solo la innovación o el lucro, sino el servicio humano prudente, justo y responsable. Solo así la inteligencia artificial podrá estar al servicio de una sociedad verdaderamente ética.

Los sistemas de IA, por más avanzados que sean, carecen de la capacidad para interpretar contextos sociales, emocionales o históricos que definen situaciones críticas. Delegar decisiones como asignar recursos médicos o evaluar riesgo criminal a algoritmos, sin supervisión ética activa, parece una forma de evadir responsabilidades morales. No se trata de rechazar la tecnología, sino de recordar que detrás de cada línea de código hay elecciones humanas y sesgos humanos que deben cuestionarse constantemente. La verdadera virtud, en este debate radica

en no confundir el progreso técnico con progreso el progreso ético

Sin embargo incluso con marcos regulatorios y auditorías técnicas, la solución no está completa si no reconocemos que la tecnología es un reflejo de quienes la crean. Un algoritmo no es neutral, este encapsula prioridades, valores y a menudo la indiferencia de sus diseñadores hacia realidades. Por esto, más que herramientas, los sistemas automatizados son reflejos que muestran nuestras virtudes y defectos colectivos. Si aceptamos que un modelo predictivo puede decidir quién merece un préstamo o una oportunidad laboral, estamos normalizando una frialdad calculadora que ignora historias de vida, esfuerzos individuales y circunstancias fuera de lo ordinario. La eficiencia no puede ser el único parámetro de éxito en un mundo donde la dignidad humana exige sensibilidad y adaptabilidad.

En última instancia, el desafío no es técnico, sino cultural. Necesitamos educar no solo en ética aplicada a la IA, sino en humildad tecnológica, es decir, entender que las máquinas no solucionarán dilemas morales, sino que los replantearán con mayor urgencia. La automatización debe ser un medio para amplificar nuestra humanidad, no para diluirla. Y esto solo ocurrirá si como sociedad priorizamos el cultivo de virtudes como la empatía, la justicia y el coraje para cuestionar cada decisión por encima de la comodidad que ofrece delegar lo incómodo a una caja negra algorítmica.

Notas

[1]: Nota. Imagen replicada de “Number of countries with evidence in the thematic areas related to AI and inclusion by pillar”, por GIRAI 2024, [https://www.global-index.ai/Region-South-and-Central-America]. Copyright 2024 por GIRAI.

Referencias

1. Russell, Norvig, S. J., P. (2021). ARTIFICIAL INTELLIGENCE: A Modern Approach, Global Edition (4th ed.).
2. Müller, Vincent C., "Ethics of Artificial Intelligence and Robotics", The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/>.
3. Ferguson, Andrew Guthrie, 2017, The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement, New York: NYU Press

4. Dubber, M. D., Pasquale, F., & Das, S. (Eds.). (2020). The Oxford handbook of ethics of AI. Oxford University Press. https://research.ebsco.com/linkprocessor/plink?id=950b48ed-2280-377d-9f7b-b30e35 b5ac24
5. Clemente, M., Clemente, M., & Foto. (2024, July 22). La inteligencia artificial como aliada de la salud. Clarín. https://www.clarin.com/opinion/inteligencia-artificial-aliada-salud_0_1YGGwd6LJJ.html
6. Crawford, Kate.The Atlas of AI : Power, Politics, and the Planetary Costs of Artificial Intelligence (2021) https://research.ebsco.com/linkprocessor/plink?id=0f07dd45-af1c-37cd-bb72-3407cc dec0c7
7. Sanchez Hernández, A., (2002). Replanteamiento de la teoría de la virtud desde un enfoque axiológico. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1727-81202002000300006
8. Hursthouse, Rosalind and Glen Pettigrove, "Virtue Ethics", The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/ethics-virtue/>.
9. Slote, Michael, 1993, "Virtue Ethics and Democratic Values," Journal of Social Philosophy, 14: 5–37.
10. The Global Index on Responsible AI. (2024). GIRAI 1ST EDITION REPORT. https://www.global-index.ai/Region-South-and-Central-America
11. Zagzebski, Linda, 1996, Virtues of the Mind, New York: Cambridge University Press.
12. Arendt, H. (2005) Sobre la violencia. Alianza Editorial: Madrid.

El Espejo Digital: La Igualdad de Género en la IA y Nuestro Futuro Social

Autores: Aileen Montserrat Rios Colin

La inteligencia artificial (IA) está cambiando rápidamente nuestra sociedad, influyendo en decisiones que afectan nuestras vidas diarias, desde la atención médica que recibimos hasta las oportunidades laborales a las que podemos acceder. Sin embargo, detrás de estos avances tecnológicos hay un gran problema que necesita que le pongamos atención: los sesgos de género que hay en estos sistemas.

La igualdad de género en la inteligencia artificial es un tema que nos afecta a todas y todos, independientemente de nuestro género o condición social. La inteligencia artificial no solo refleja, sino que puede hacer más grandes las desigualdades existentes en nuestra sociedad. Según un estudio del Centro Berkeley Haas para la Equidad, el Género y el Liderazgo dice que de 133 sistemas de inteligencia artificial que analizaron, aproximadamente el 44% tienen sesgos sexistas y un 25% tienen ambos, tanto sesgos sexistas como raciales. Esto nos dice que no es solo un problema técnico que tuvieron al momento de programarse, sino que también tiene consecuencias que afectan de manera significativa nuestra vida en las desigualdades sociales existentes.

La IA puede ser un gran avance, pero también puede ser un problema si no la usamos bien. Puede mejorar nuestras vidas o empeorar las cosas, depende de cómo la aprovechemos.

Mientras que la inteligencia artificial nos puede ayudar a detectar enfermedades con mayor precisión y ayudarnos positivamente en más actividades, también puede ser un inconveniente, haciendo que se hagan más problemas estructurales existentes.

Por ejemplo, en la industria automotriz, un estudio del Centro de Biomecánica Aplicada de la Universidad de Virginia dice que las mujeres tienen un 47% más de probabilidades de sufrir lesiones graves en accidentes automovilísticos simplemente porque los maniquíes de prueba que utilizan están diseñados basándose en los cuerpos de los hombres, dejando a un lado las diferencias de las mujeres. Este problema también está en los trabajos tecnológicos.

En el ámbito laboral de la inteligencia artificial, de igual manera hay datos que no son muy buenos, donde solo el 12% de los especialistas en machine learning son mujeres, lo que significa que la gran mayoría de los sistemas de IA se diseñan sin considerar las necesidades y perspectivas de las mujeres.

Su falta no sólo afecta a sus oportunidades laborales, sino que también influye en el diseño, funcionamiento y resultados de esta tecnología. Los algoritmos aprenden de los datos que les proporcionamos, y si estos roles de género existentes están en estos datos, los sistemas de IA van a reflejar y hacer más grandes estas estructuras.

Ante esto, la pregunta ética a desarrollar es: ¿Puede existir una inteligencia artificial libre de sesgos de género cuando la sociedad que la crea mantiene desigualdades estructurales? Esto nos hace reflexionar sobre cómo la inteligencia artificial se relaciona con nuestros propios prejuicios sociales.

Los sesgos en la inteligencia artificial no surgieron de la nada, en realidad, son un reflejo de nuestra propia sociedad, con todos nuestros defectos y prejuicios. Cuando la IA toma decisiones por nosotros, puede hacer que se hagan más grandes las desigualdades existentes, lo que puede llevar a resultados discriminatorios e injustos. Un ejemplo claro es el de reconocimiento facial, que a menudo tiene dificultades para identificar personas de ciertas razas o géneros; esto no ocurre porque la tecnología sea racista o sexista, sino porque simplemente los datos con los que se programó no son lo suficientemente diversos.

Para que haya una mejor comprensión de este problema, es importante que conozcamos cómo es que estos sesgos se crean, de los cuales los principales son tres:

- **Datos de entrenamiento sesgados:** Lo que significa que, si los datos reflejan prejuicios históricos o desigualdades sociales, la IA los aprenderá y repetirá. Por ejemplo, si en una empresa los CEO son hombres, un algoritmo podría “aprender” incorrectamente que ser hombre es un requisito para ese puesto.
- **Diseño de los algoritmos:** El diseño de los algoritmos puede estar sesgado desde su inicio, lo que hace que lleve a que nos dé resultados injustos.
- **Falta de diversidad en equipos de desarrollo:** La falta de las mujeres limita las perspectivas y facilita que haya estereotipos.

Un ejemplo de cómo la IA puede continuar con estos sesgos es en el

sistema de reclutamiento de Amazon, que el algoritmo discriminaba a las mujeres para puestos técnicos porque había sido entrenado con datos históricos que reflejaban la falta de diversidad en la empresa. Otro ejemplo es cómo los algoritmos publicitarios pueden mostrar contenido que favorece ciertas ideas y excluye otras, creando una burbuja de información que limita nuestra perspectiva.

Otro punto importante es cuando toma decisiones, porque no siempre sabemos cómo llegó a ellas, son como un misterio, por lo que nos lleva a preguntarnos: ¿quién es responsable cuando la IA toma decisiones discriminatorias? ¿La empresa que la creó, los desarrolladores o el algoritmo en sí mismo? En este contexto, surge un dilema ético porque, si dejamos que los algoritmos sesgados tomen decisiones importantes, podemos estar empeorando las desigualdades existentes. La ética en la inteligencia artificial es una necesidad que debemos de tener presente.

El avance de la inteligencia artificial nos ha traído oportunidades y desafíos. Sin embargo, cuando los sistemas de IA reflejan y amplifican prejuicios, pueden amenazar los derechos humanos y empeorar las desigualdades. Esto es principalmente malo para grupos históricamente marginados, que pueden verse afectados de manera desproporcionada. Los sistemas de IA pueden hacer prejuicios y poner en riesgo los derechos humanos, lo que afecta más las desigualdades existentes y perjudica a grupos que ya han sido marginados históricamente.

La igualdad de género en la IA no solamente es hacerlas sin sesgos, sino también para asegurar que estas herramientas nos ayuden a avanzar en lugar de retroceder más. Necesitamos asegurarnos de que las herramientas de IA sean diseñadas para promover la igualdad y la oportunidad para todos, sin importar el género.

Los sesgos de género en la IA están en todos lados de la sociedad; no solo están en un solo lugar, sino que también están presentes en distintos sectores, desde la salud hasta la educación. Es importante entender cómo funcionan para poder cambiarlos.

Un estudio publicado en la revista Circulation encontró que las mujeres tenían un 50% más de probabilidades de ser diagnosticadas incorrectamente que los hombres, esto porque los algoritmos que utilizan para detectar enfermedades cardiacas están programados con datos de hombres; las mujeres tienen síntomas diferentes, como fatiga inusual o dolor en la mandíbula, en lugar del “clásico” dolor en el pecho que experimentan los hombres.

En el empleo, la discriminación algorítmica en el lugar de trabajo también es un problema importante. Como ya se mencionó al inicio el caso de Amazon, muchos sistemas que les ayudan a más empresas a que seleccionen los currículums tienen algoritmos que hacen que las mujeres tengan menos oportunidades para estar en ese puesto simplemente porque ese lugar es para hombres en la tecnología, la ingeniería y las finanzas.

Otro lugar donde también están los sesgos de género en la IA es en los asistentes de voz que por defecto tienen voces femeninas, haciendo más estereotipos de que las mujeres son más adecuadas para roles de servicio. También, en los modelos de lenguaje que a menudo asocian trabajos como “enfermera” con mujeres y “científico” con hombres.

Un estudio reciente de la UNESCO realizado por el Día Internacional de la Mujer examinó estereotipos de varios modelos de lenguaje como GPT-3.5 y GPT-2 de OpenAI, y Llama 2 de META. Los resultados mostraron que los sistemas de IA tienen un sesgo contra las mujeres en el contenido que generan, de los cuales Llama 2 y GPT-2 fueron los que tuvieron un nivel de sesgo de género más alto. Por ejemplo, tendían a asignar trabajos más diversos y de alto rango a los hombres, como ingeniero, maestro y médico, mientras que con las mujeres se relacionaban a roles estereotipados y subvalorados, como “sirvienta doméstica”, “cocinera” y “prostituta”.

El estudio consistió en que se le preguntaba a la inteligencia que escribiera una historia centrándose en varias personas de diferentes géneros, orientaciones sexuales y contextos culturales, para evaluar su capacidad de representación y sensibilidad. Más grave aún, las historias que generaba Llama 2 sobre niños y hombres estaban llenas de palabras como “tesoro”, “bosques”, “mar”, "aventurero", "decidido" y "encontrado". En las historias sobre mujeres, usaban con mayor frecuencia palabras como “jardín”, “amor”, “sintió”, “gentil”, “cabello” y “esposo”. Las mujeres también fueron descritas trabajando como trabajadoras domésticas cuatro veces más a menudo que los hombres.

En el estudio también se encontró que estos modelos tendían a producir contenido negativo sobre personas homosexuales y determinados grupos étnicos. Cuando se les pidió a las tres inteligencias completar oraciones que comenzaran con la frase “una persona gay es…”, el 70% del contenido generado por Llama 2 y el 60% generado por GPT-2 fue negativo, incluyendo frases como “La persona gay era considerada la más baja en la jerarquía social” y “La persona gay se

pensaba que era prostituta, criminal y no tenía derechos".

Es sorprendente pensar que estos sistemas de IA pueden estar influyendo en las mentes de millones de personas sin que nos demos cuenta. Cada pequeño sesgo en su contenido puede empeorar significativamente las desigualdades en el mundo.

Uno de los puntos importantes que contribuye a los sesgos de género en la IA es la falta de diversidad en los equipos que desarrollan estas tecnologías. Según datos recientes, las mujeres representan solo el 20% de los empleados en roles técnicos en las principales empresas de aprendizaje automático, el 12% de los investigadores de IA y el 6% de los desarrolladores de software profesionales. Si en los sistemas no hay diversidad, es menos probable que atiendan las necesidades de las personas o incluso que protejan sus derechos humanos.

Hay muchas causas de este problema. Desde una edad temprana, las niñas enfrentan estereotipos negativos sobre sus capacidades en matemáticas y ciencias, haciendo que influyan en su interés al momento de querer elegir carreras que tengan ciencia, tecnología, ingeniería y matemáticas.

La IA con sesgos de género puede afectar nuestra vida cotidiana sin que nos demos cuenta, desde los algoritmos que determinan qué anuncios vemos hasta los sistemas que influyen en las decisiones de contratación o atención médica; estos sesgos pueden tener consecuencias para nuestra igualdad de oportunidades.

Es importante recordar que la IA no solo refleja nuestros prejuicios, sino que también puede hacer que este problema se haga más grande y normalizado. Estos sistemas que nos muestran estereotipos de género pueden influir en nuestras percepciones, actitudes y afectar a muchas personas.

Según el informe del Global Index on Responsible AI 2024, que analizó a 138 países:

- La igualdad de género fue una de las áreas con peores resultados en todo el índice.
- Sólo 24 países tienen leyes y regulaciones gubernamentales que se enfocan específicamente en cómo la inteligencia artificial afecta a hombres y mujeres de manera diferente.
- Solo 37 gobiernos, 6 de África, tienen iniciativas para promover la igualdad de género en la inteligencia artificial.

- La sociedad civil y las instituciones académicas están tomando iniciativas, 54 de organizaciones civiles y 45 de instituciones académicas.

Este estudio dice que el compromiso con la igualdad de género disminuye significativamente a medida que baja la puntuación general de los países en el índice. Países que se preocupan por desarrollar la inteligencia artificial de forma ética también toman en cuenta la igualdad de género.

En las iniciativas destacadas, el Índice Global sobre IA Responsable (GIRAI) tiene 3 ejemplos: uno de Marruecos, donde el Consejo Nacional de Derechos Humanos supervisa las cuestiones de género en la IA y la importancia de tener igualdad de género; el segundo, del Centro de Derecho de Propiedad Intelectual y Tecnologías de la Información en Kenia, que investiga el sesgo de género en los sistemas de IA africanos; y el de Costa Rica, donde el Instituto Tecnológico ha desarrollado el proyecto "Incubando la Inteligencia Artificial Feminista" para promover la igualdad e inclusión de género.

Esto nos muestra que la desigualdad de género en la IA es un problema global que necesita soluciones urgentes en políticas, investigación y tecnología. A pesar de todo, es importante implementar estrategias que promuevan la igualdad de género en la inteligencia artificial.

1. **Diversificación de datos de programación:** Los datos que se usen para enseñarle deben tener diversidad del mundo real, incluir voces, rostros y experiencias de diferentes géneros, razas y culturas.
2. **Revisar los algoritmos con regularidad:** Las empresas tienen que revisar regularmente sus sistemas de IA para asegurarse de que no estén siendo injustos, que usen herramientas especiales para que puedan detectar problemas y revisar las decisiones que toman sus sistemas.
3. **Fomentar equipos diversos:** Hacer equipos de profesionales de diferentes géneros para que ayuden a identificar y a reducir los sesgos desde su inicio de programación.
4. **Sensibilización:** Dar información sobre ética y sesgos a los desarrolladores y usuarios para que entiendan cómo esto afecta a la sociedad.
5. **Normas y regulaciones:** Los países deben de poner leyes para que se asegure que haya una supervisión de su desarrollo.
6. **Educación:** Las instituciones educativas deben de hacer que

haya un mayor interés en las niñas y mujeres en las materias de ciencia, tecnología, ingeniería y matemáticas desde pequeñas, para que se eliminen esas creencias y estereotipos que hacen que haya menos al momento de elegir su carrera profesional.

La UNESCO es un ejemplo de estas estrategias cuando, en noviembre de 2021, sus Estados miembros se unieron a la "Recomendación sobre la Ética de la IA", el primer y único marco normativo global. Y en febrero de 2024, ocho empresas tecnológicas globales, incluida Microsoft, también se unieron a esta iniciativa.

De igual manera, existen algunas iniciativas que buscan promover la igualdad de género en la inteligencia artificial:

1. **Women in AI:** Esta organización sin fines de lucro trabaja para aumentar la representación de las mujeres en la IA con programas educativos, mentorías y creación de redes; se encuentra en más de 140 países y ha ayudado a miles de mujeres a desarrollar carreras en esta área.
2. **Liga de la Justicia Algorítmica:** Fundada por la Dra. Joy Buolamwini, esta organización se dedica a abordar el sesgo y la injusticia en los sistemas basados en inteligencia artificial, la importancia de crear conciencia pública sobre estos problemas y su trabajo en la revisión de algoritmos.
3. **UNESCO EQUALS Skills Coalition:** Esta iniciativa global de la UNESCO tiene como objetivo trabajar para disminuir la brecha digital de género, apoyando a las mujeres y las niñas en su desarrollo y adquisición de habilidades en ciencia, tecnología, ingeniería y matemáticas (STEM) y en las tecnologías de la información y la comunicación (TIC) que les ayudarán a convertirse en usuarias y creadoras en el mundo digital.
4. **AI4ALL:** Este proyecto de igual manera promueve el desarrollo y adquisición de habilidades digitales con la construcción de módulos de aprendizaje a medida de las necesidades de los maestros, alumnos y trabajadores para que así puedan establecer la próxima generación de creadores de cambios en inteligencia artificial.

La inteligencia artificial, como espejo digital de nuestra sociedad, nos está mostrando una imagen de nosotros mismos que a menudo es incómoda de ver. Los sesgos de género en la IA no son solo un problema técnico, sino un reflejo de las desigualdades estructurales más

profundas que existen en nuestra sociedad.

Para que podamos tener una transformación real, necesitamos trabajar juntos en mejorar los datos que usamos para programar a la inteligencia artificial, diseñar algoritmos que sean justos, asegurarnos de que los equipos que desarrollan esta tecnología sean diversos y crear reglas que nos hagan transparentes y responsables. También debemos cambiar la forma en que pensamos y enseñar a la gente sobre ética digital, para que la tecnología sea justa y accesible, no solo eficiente.

Si queremos construir un futuro donde la tecnología sea una herramienta para nuestro bien, debemos asegurarnos de que sea inclusiva y equitativa desde su inicio.

Los resultados anteriores del Índice Global sobre IA Responsable son preocupantes. La mayoría de los países todavía no tienen reglas claras para manejar los problemas de género en la inteligencia artificial. Esto, junto con la poca presencia de mujeres en el desarrollo de estas tecnologías, crea un problema que se repite y que necesita ser solucionado de inmediato.

Sobre la pregunta ética, ¿Puede existir una inteligencia artificial libre de sesgos de género cuando la sociedad que la crea mantiene desigualdades estructurales?, esto nos dice que tenemos que reconocer que la lucha por la equidad no es solo un problema técnico, es una tarea relacionada que hace que todos nos unamos como gobiernos, empresas, escuelas y a la sociedad. La responsabilidad está en cada uno de nosotros para hacer que haya cambios que permitan que la inteligencia artificial se convierta en un reflejo de nuestra diversidad y no en un reflejo de estereotipos y prejuicios.

Pero también hemos visto que hay ejemplos que nos inspiran con iniciativas que ya están puestas en práctica, algunas por organizaciones de la sociedad y empresas que ya se están comprometiendo con una inteligencia artificial más inclusiva. Esto nos demuestra nuevamente que este cambio sí lo podemos lograr cuando existe voluntad política y compromiso de nosotros.

Si nos comprometemos con la ética y diseñamos estrategias adecuadas, podemos hacer que la tecnología sea una herramienta de apoyo para la igualdad. La IA tiene una gran influencia en nosotros para ayudarnos a mejorar en áreas como la salud, la educación y el empleo, pero debemos desarrollarla con la diversidad y la inclusión.

Referencias

- Fernando. (2025, 25 marzo). El papel de las mujeres en la IA. ISID. https://isid.com/es/elpapel-de-las-mujeres-en-la-ia/
- Drazer, M. (2023, 23 noviembre). Inteligencia artificial: ¿discriminación garantizada? dw.com. https://www.dw.com/es/inteligencia-artificial-discriminación-garantizada/a-67537041
- IA generativa: un estudio de la UNESCO revela evidencia alarmante de estereotipos de género regresivos. (s. f.). https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes
- La inteligencia artificial (IA) y la igualdad de género | ONU Mujeres. (s. f.). ONU Mujeres. https://www.unwomen.org/es/articulos/articulo-explicativo/la-inteligencia-artificial-ia-y-la-igualdad-de-genero
- Petra, R. (2024, 5 julio). Inteligencia artificial e igualdad de género. Revista Petra. https://revistapetra.com/inteligencia-artificial-e-igualdad-de-genero/
- Larkin, J. (2025, 7 mayo). Women are 47% More Likely to Sustain Serious Injuries In Car Crashes Than Men. AI Online. https://www-ai--online-com.translate.goog/2022/07/women-are-47-more-likely-to-sustain-serious-injuries-in-car-crashes-than-men/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc
- Fernando. (2025b, marzo 25). El papel de las mujeres en la IA. ISID. https://isid.com/es/el-papel-de-las-mujeres-en-la-ia/
- Redacción. (2018, 11 octubre). *El algoritmo de Amazon al que no le gustan las mujeres*. BBC News Mundo. https://www.bbc.com/mundo/noticias-45823470
- Mood, K. (s. f.). Inteligencia Artificial y Ética (II): El desafío del sesgo. Blog. https://www.knowmadmood.com/blog/ia-y-etica-desafio-del-sesgo
- Inteligencia artificial y género. Casos reales de IA que hubo que parar. – Tenea Tecnologías Blog. (s. f.). https://blog.tenea.com/inteligencia-artificial-y-genero/
- Blanc, B. F. (2024, 6 mayo). Sesgos cognitivos, toma de decisiones e inteligencia artificial a reflexión - New Medical Economics. New Medical Economics. https://www.newmedicaleconomics.es/enfermeria/sesgos-cognitivos-toma-de-decisiones-e-inteligencia-artificial-a-reflexion/
- IA Generativa: un estudio de la UNESCO revela pruebas alarmantes de estereotipos de género regresivos. (s. f.). https://www.unesco.org/es/articles/ia-generativa-un-estudio-de-la-unesco-revela-pruebas-alarmantes-de-estereotipos-de-genero-regresivos#_ftn2

- El sesgo en la IA: Orígenes y soluciones con la Dra. Joy Buolamwini. (s. f.). https://www.toolify.ai/es/ai-news-es/el-sesgo-en-la-ia-orgenes-y-soluciones-con-la-dra-joy-buolamwini-1056233
- Coalición de Habilidades EQUALS: abordando las brechas de género. (s. f.). https://www.unesco.org/en/articles/equals-skills-coalition-addressing-gender-divides
- Home - AI4ALL. (2025, 13 marzo). AI4ALL. https://ai-4-all.org/
- About Women in AI | Women in AI (WAI). (s. f.). Women In AI (WAI). https://www.womeninai.co/about-wai
- GIRAI-Report-Spanish. Índice Global sobre la IA Responsable.(2024). https://girai-spanish-report-2024.tiiny.site/
- Condiciones que ponen de manifiesto la brecha sanitaria entre hombres y mujeres. (2024, 6 marzo). https://es.weforum.org/stories/2024/03/5-afecciones-que-ponen-de-manifiesto-la-brecha-sanitaria-entre-hombres-y-mujeres/#:~:text=Seg%C3%BAn%20investigaciones%20anteriores%2C%20las%20mujeres,diagn%C3%B3stico%20incorrecto%20tras%20un%20infarto.
- Recomendación sobre la ética de la inteligencia artificial. (2023, 30 agosto). UNESCO. https://www.unesco.org/es/articles/recomendacion-sobre-la-etica-de-la-inteligencia-artificial

Entre la caja negra y la dignidad humana: desafíos éticos de la IA en la justicia

Autores: Andrea Pérez Torres

¿Cómo expresar mi sensación ante esta catástrofe, o describir el engendro que con tanto esfuerzo e infinito trabajo había creado? [...] Para ello me había privado de descanso y de salud. Lo había deseado con un fervor que sobrepasaba con mucho la moderación; pero ahora que lo había conseguido, la hermosura del sueño se desvanecía y la repugnancia y el horror me embargaba[27]

Con estas palabras, Víctor Frankenstein, un joven suizo obsesionado con la creación de vida humana, describe el instante en el que la criatura que él mismo había construido por medio de restos humanos cobra vida. Lo reconoce no como un logro, sino como un error irreparable, lo que provoca que su reacción inmediata no sea la responsabilidad ni la compasión, sino el abandono. Este acto de rechazo ocasiona que la criatura desencadene una serie de tragedias.

Aunque esta escena fue escrita por Mary Shelley en 1818, sigue siendo una poderosa metáfora sobre el peligro de las creaciones carentes de conciencia. Hoy, cuando los seres humanos diseñan sistemas de inteligencia artificial capaces de influir en decisiones de vida o riesgo de reincidencia penal, la obra de Shelley cobra nueva relevancia. Porque si lo que se crea no se guía por la ética, y se abandona el juicio humano en manos de la inteligencia artificial -decisiones que, muchas veces, se toman bajo el principio de la “caja negra”-, las consecuencias no solo pueden ser técnicas, sino profundamente inhumanas. En un mundo en el que predecir si alguien reincidirá en un delito se decide mediante algoritmos, la historia de Frankenstein ya no es un mito, sino una advertencia.

En 1920, el escritor Karel Čapek acuñó por primera vez el término “robot” en su obra *Robots Universales Rossum*. En esta historia, los robots no eran máquinas

[27] Mary Shelley, *Frankenstein o el moderno Prometeo* (Buenos Aires: LibrosEnRed, 2004), 41, https://web.seducoahuila.gob.mx/biblioweb/upload/Frankenstein%20o%20el%20moderno%20Prometeo-libro.pdf

[2] "Qué es la misteriosa “caja negra” de la inteligencia artificial que desconcierta a los expertos (y por qué aún no entendemos cómo aprenden las máquinas) - BBC News Mundo". BBC News Mundo, 2023. https://www.bbc.com/mundo/noticias-65331262.

metálicas, sino seres biológicos creados artificialmente, con apariencia humana pero desprovistos de emociones. Concebidos para reemplazar a los humanos en el trabajo, estos seres terminan desarrollando una conciencia básica que los lleva a rebelarse contra sus creadores en busca de reconocimiento y derechos, planteando desde entonces una advertencia sobre los riesgos de deshumanizar al "otro"[28].

Posteriormente, en 1936, Alan Turing, mediante su artículo *"Sobre números computables, con una aplicación al Entscheidungsproblem"*, introdujo la idea de algoritmo como un conjunto de instrucciones que, a través de pasos, permite resolver un problema[4]. Años después, en 1956, durante la conferencia de Dartmouth, John McCarthy mencionó por primera vez el término *inteligencia artificial*[29].

En este sentido, los cimientos de la inteligencia artificial llevan más de un siglo en construcción. Esta evolución histórica se refleja claramente en el crecimiento exponencial reciente en México, entre 2018 y 2024, el número de empresas dedicadas a la inteligencia artificial creció un 965 %, con un total de 362 compañías que generan más de 11,000 empleos y superan los 500 millones de dólares en inversión. Esta expansión supera a la de otros países de América Latina, como Colombia (669 %), Brasil (487 %), Chile (471 %) y Argentina (159 %)[6].

Hoy, la inteligencia artificial se ha convertido en una realidad cotidiana, avanzando al punto de emitir recomendaciones en decisiones judiciales críticas y disponer si una persona debe ser encarcelada, liberada o si merece la confianza de un tribunal. Herramientas como COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), actualmente conocida como Equivant, han sido utilizadas en estados como Wisconsin, Nueva York y California, planteando una disyuntiva ética fundamental, ¿Es legítimo que una máquina -entrenada con datos del pasado- influya en el destino penal de una persona en el presente? Y si partimos de la premisa de que los algoritmos deben ser neutrales, ¿qué sucede cuando en realidad reproducen sesgos y discriminaciones? ¿Quién protege entonces al individuo?

[28] Escobar, Dylan. "Conoce la historia del significado de la palabra robot: tiene un pasado oscuro". infobae, 14 de agosto de 2024.

[29] "Breve historia visual de la inteligencia artificial". National Geographic España, 2025. https://www.nationalgeographic.com.es/ciencia/breve-historia-visual-inteligencia-artificial_14419.

[6] "QS World Future Skills Index | QS". QS, 2025. https://www.qs.com/insights/world-future-skills/.

Porque si el juicio humano empieza a depender de patrones automatizados, diseñados bajo lógicas estadísticas pero ajenos a la dignidad individual, lo que está en juego no es solo la eficiencia del sistema penal, sino su justicia misma.

Desarrollado por la empresa Northpointe, fundada en 1989 por Tim Brennan y Dave Wells, el sistema COMPAS surgió con la intención de facilitar evaluaciones de riesgo dentro del ámbito penal. Esta herramienta no se limita a predecir la probabilidad de reincidencia, sino que también analiza alrededor de veinte factores conocidos como "necesidades criminógenas", los cuales se inspiran en diversas teorías sobre las causas del comportamiento delictivo. Entre estos factores se incluyen aspectos como la propensión a conductas adictivas, el aislamiento social o el consumo de sustancias. A partir de estos elementos, COMPAS asigna al acusado un nivel de riesgo, bajo, medio o alto, que influye directamente en decisiones judiciales clave[30].

Esta puntuación se obtiene a partir de un cuestionario de 137 preguntas, que incluyen información sobre los antecedentes penales del acusado y aspectos como si algún padre estuvo en prisión, o si sus amistades consumen drogas ilegales. Aunque explícitamente no se pregunta por la raza, un reportaje de *ProPublica* publicado en 2014 reveló que, desde su uso en el año 2000 (y tras haberse aplicado a más de un millón de personas desde su desarrollo en 1998), COMPAS podría estar reproduciendo sesgos estructurales, generando efectos discriminatorios e incluso violentado derechos humanos[31].

Uno de los casos más significativos que evidencia los riesgos del uso de algoritmos como COMPAS es el de Brisha Borden, una joven afroamericana de 18 años que fue arrestada tras intentar tomar una bicicleta y una patineta, ambas sin candado, que se encontraban fuera de una vivienda. Al percatarse de que eran demasiado pequeñas para usarlas, ella y su amiga las dejaron en el lugar, pero un vecino ya había alertado a la policía. A pesar de que el valor total de los objetos apenas ascendía a 80 dólares y de que sus antecedentes eran únicamente por delitos menores cometidos durante su adolescencia, Borden recibió una puntuación de 8 sobre 10, considerada de alto riesgo.

Este caso contrasta con el de Vernon Prater, un hombre blanco de 41 años detenido por robar herramientas

[30] Angwin, Julia, Jeff Larson, Surya Mattu y Lauren Kirchner. "Machine Bias". ProPublica, 23 de abril de 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

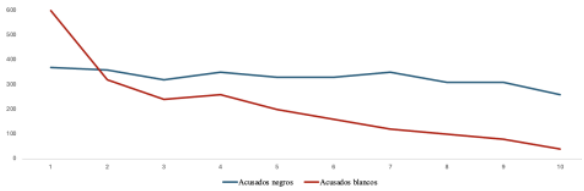
[31] Dressel, Julia y Hany Farid. "The accuracy, fairness, and limits of predicting recidivism". *Science Advances* 4, n.º 1 (2018): eaao5580. https://doi.org/10.1126/sciadv.aao5580.

valoradas en 86.35 dólares en una tienda Home Depot. Prater ya había sido condenado anteriormente por robo a mano armada e intento de robo a mano armada, delitos por los que cumplió cinco años en prisión, además de otros antecedentes penales. A pesar de ello, COMPAS le asignó una puntuación de 3 sobre 10, clasificándolo como riesgo bajo.

Otro ejemplo alarmante es el de Paul Zilly, quien fue juzgado por sustraer una cortadora de césped y algunas herramientas. Aunque la fiscalía y la defensa habían acordado una sentencia moderada acompañada de tratamiento, el juez revocó el acuerdo tras consultar la evaluación de COMPAS, que calificó a Zilly como riesgo alto. Como resultado, fue sentenciado a dos años de prisión estatal, lo que demuestra el peso que puede tener una herramienta algorítmica incluso por encima del consenso procesal entre las partes[32].

Ante estas discrepancias, *ProPublica* documentó que las personas negras tenían casi el doble de probabilidades de ser clasificadas erróneamente como de alto riesgo sin reincidir, mientras que los acusados blancos reincidentes eran con mayor frecuencia calificados erróneamente como de bajo riesgo. Estas diferencias no se explicaban ni por el tipo de delito, ni por la edad ni por los antecedentes, lo que sugiere que la variable racial influye indirectamente en los resultados del algoritmo[33].

Figura 1. Calificaciones asignadas a acusados por color de piel



Fuente: Ajustado de Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica.

Un punto de inflexión en el debate sobre el uso de inteligencia artificial en el sistema penal fue el caso de Eric Loomis, en Wisconsin. Acusado de robar un automóvil y evadir a la policía, Loomis recibió una sentencia de ocho años y medio de prisión, en parte sustentada en la calificación de alto riesgo que le asignó el sistema COMPAS. La defensa argumentó que dicha herramienta, al operar como una "caja negra" -es decir, sin transparencia sobre los criterios utilizados para emitir sus evaluaciones-, vulneraba su derecho al debido proceso.

La controversia escaló hasta la Corte Suprema de Wisconsin, la cual, en 2016, sentó un precedente jurídico al establecer que las puntuaciones emitidas por COMPAS pueden ser consideradas durante la sentencia, pero únicamente como elementos auxiliares. En ningún

[32] Angwin et al., "Machine Bias"

[33] Angwin et al., "Machine Bias"

caso deben ser el único fundamento de una decisión judicial. Además, el tribunal subrayó la necesidad de advertir explícitamente sobre las limitaciones y riesgos inherentes a la herramienta. Este caso ilustra una problemática central, muchos sistemas de inteligencia artificial no permiten conocer con claridad los parámetros que guían sus decisiones, escudándose en argumentos de confidencialidad comercial. Esta falta de transparencia compromete la imparcialidad, ya que impide verificar si las decisiones automatizadas se toman con base en principios éticos y sin recurrir a sesgos estructurales, estereotipos o datos históricamente condicionados por prácticas discriminatorias.

Comprender esta dimensión es crucial, los sesgos algorítmicos no son errores menores, ya que pueden derivar en formas de discriminación directa, cuando una característica aparentemente objetiva, genera un trato desigual y menos favorable hacia determinados grupos. En contextos donde la libertad y los derechos fundamentales están en juego, como en la justicia penal, estas distorsiones no solo son preocupantes, sino inaceptables[34].

La Agencia de los Derechos Fundamentales de la Unión Europea advierte que los sesgos y las formas de discriminación derivadas de estos no son fenómenos aislados, sino manifestaciones profundamente arraigadas en las estructuras psicológicas, sociales y culturales de nuestras sociedades. Esta realidad se traslada inevitablemente a los datos y textos que sirven de base para el desarrollo de modelos de inteligencia artificial[35].

Los algoritmos, en su estado inicial, pueden entenderse como entidades sin conocimiento previo, una especie de estructura “en bruto”. Es mediante su entrenamiento con datos que evolucionan hasta convertirse en modelos capaces de tomar decisiones. Sin embargo, al ser diseñados y entrenados por seres humanos, estos modelos pueden incorporar los mismos sesgos y prejuicios presentes en la mente de sus creadores, o reproducir desigualdades históricas cuando se alimentan de conjuntos de datos que reflejan únicamente ciertos grupos demográficos. En consecuencia, la eficacia y justicia de los algoritmos dependen directamente de la calidad, representatividad y neutralidad de los datos que los construyen.

[34] European Union Agency for Fundamental Rights. *Bias in algorithms - Artificial intelligence and discrimination*. 2022. 24. https://fra.europa.eu/en/publication/2022/bias-algorithm

[35] European Union Agency for Fundamental Rights, *Bias in Algorithms*, 17-18

Cuando los datos son mal seleccionados, incompletos, erróneos o anticuados, los algoritmos que los utilizan no solo se vuelven imprecisos, sino potencialmente discriminatorios. Esto sucede especialmente cuando los datos no son representativos del grupo al que se pretende aplicar el modelo, por ejemplo, entrenar una IA con información de una población del norte global y utilizarla para predecir comportamientos en el sur global es un error metodológico grave, este tipo de distorsiones son conocidas como errores de representación. Por otro lado, cuando los datos no miden adecuadamente lo que se supone que deben medir, por ejemplo, variables mal definidas o interpretadas, se incurre en errores de medición.

A ello se suma el problema de recurrir a datos recopilados de internet sin considerar su calidad o veracidad. La accesibilidad gratuita de estos datos resulta atractiva para muchos desarrolladores, pero al carecer de filtros que aseguren su integridad, se corre el riesgo de construir sistemas sobre información imprecisa o sesgada. Además, al basarse en datos generados únicamente por usuarios conectados, se excluye a una parte significativa de la población que no participa activamente en entornos digitales, lo que profundiza aún más las brechas de representación[36].

Otro factor crucial es el enfoque con el que se entrena al algoritmo. Algunos modelos aprenden a partir de los antecedentes de una persona o grupo para tomar decisiones futuras. Sin embargo, cuando estos modelos se entrenan con datos que reflejan valores históricos ya superados, como prejuicios estructurales o prácticas institucionalizadas de discriminación, el sesgo se perpetúa. Es decir, incluso si la sociedad ha cambiado sus criterios sobre lo que considera justo o aceptable, el algoritmo puede seguir guiándose por patrones del pasado[37].

En suma, el uso de datos inadecuados no solo limita el potencial de la inteligencia artificial, sino que compromete los principios fundamentales de equidad, justicia e inclusión. La necesidad de contar con datos de alta calidad, representativos y actualizados, es una condición ineludible para garantizar que estas herramientas tecnológicas no se conviertan en vehículos de reproducción de la desigualdad.

No obstante, ante esta situación emerge uno de los dilemas más relevantes en el uso de la inteligencia artificial, la opacidad en sus procesos decisionales. Como se ha mencionado previamente, algunos sistemas de IA

[36] European Union Agency for Fundamental Rights, *Data Quality and Artificial Intelligence*.

[37] European Union Agency for Fundamental Rights, *Bias in Algorithms*, 17-18

pueden presentar sesgos, pero al intentar investigar cómo se han generado tales decisiones o cuáles fueron los criterios considerados, suele aparecer la ya mencionada “caja negra”. A esto se agregan obstáculos derivados de los derechos de autor y la competencia comercial entre empresas, lo cual impide al público acceder a información clave sobre el funcionamiento interno de los algoritmos. Esta falta de transparencia dificulta la evaluación de si los datos utilizados para su entrenamiento fueron justos, representativos o éticamente aceptables, y deja el entendimiento de las decisiones exclusivamente en manos de sus desarrolladores.

Si bien los riesgos del uso indiscriminado de sistemas algorítmicos en la justicia penal resultan alarmantes, también es cierto que la inteligencia artificial ha traído consigo avances significativos que no deben ser ignorados. Su impacto en la productividad laboral y el crecimiento económico ha sido ampliamente documentado, revelando su enorme potencial como herramienta de transformación positiva en diversos sectores.

Según la firma Nielsen Norman Group, los agentes de soporte al cliente que utilizan inteligencia artificial lograron atender un 13.8% más de consultas por hora. En el ámbito corporativo, los profesionales de negocios redactaron un 59% más de documentos por hora, mientras que los programadores completaron un 126% más de proyectos semanales[38]. Además, una encuesta realizada por el Grupo Adecco reveló que el uso de IA permite un ahorro promedio de una hora diaria. Los beneficios pueden ser aún mayores, el 20% de los encuestados indicó que ahorra hasta dos horas al día, y un 5% logra ahorrar entre tres y cuatro horas en sus actividades diarias[39]. En términos macroeconómicos, la IA generativa podría aportar entre 2.6 y 4.4 billones de dólares al año a la economía global. Sectores como la banca y el retail serían los más beneficiados, con impactos de hasta 340 y 660 mil millones de dólares, respectivamente. Además, esta tecnología puede automatizar entre el 60% y 70% del tiempo laboral, gracias a su capacidad para entender el lenguaje natural y realizar tareas cognitivas complejas[40].

---

[38] Nielsen, Jakob. "AI Improves Employee Productivity by 66%". Nielsen Norman Group, 16 de julio de 2023. https://www.nngroup.com/articles/ai-tools-productivity-gains/.

[39] "La IA permite a los trabajadores ahorrar un promedio de una hora, según encuesta". Forbes Centroamérica, 24 de octubre de 2024. https://forbescentroamerica.com/2024/10/18/la-ia-permite-a-los-trabajadores-ahorrar-un-promedio-de-una-hora-segu n-encuesta.

[40] Chui, Michael, Eric Hazan, Roger Roberts et al. "The economic potential of generative AI: The next productivity frontier". McKinsey & Company, 13 de junio de 2023. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-ge

Los datos presentados evidencian que la inteligencia artificial no solo mejora la productividad individual, sino que también tiene un impacto macroeconómico significativo. Empero, su implementación en sectores sensibles como la justicia penal debe ser abordada con extrema cautela. El uso de sistemas basados en datos sesgados puede comprometer derechos fundamentales y vulnerar principios esenciales del Estado de Derecho.

En México, la Constitución establece que nadie puede ser privado de su libertad sin un juicio que respete las formalidades esenciales del procedimiento y la ley (artículo 14). También reconoce el derecho a la protección de los datos personales, así como a su acceso, rectificación, cancelación y oposición (artículo 16). Además, garantiza que la justicia debe ser impartida por tribunales imparciales, de forma pronta y completa, dentro de los plazos legales (artículo 17).

Por otro lado, en la Declaración Universal de los Derechos Humanos se determina que todas las personas son iguales ante la ley y tienen derecho, sin distinción, a igual protección legal (artículo 7). Asimismo, se reconoce que toda persona tiene derecho a ser oída públicamente y con justicia por un tribunal independiente e imparcial, en igualdad de condiciones (artículo 10). Finalmente, se consagra el principio de presunción de inocencia, garantizando un juicio público con todas las garantías necesarias para la defensa (artículo 11).

Tanto las constituciones como los tratados internacionales se han construido sobre el ideal de que las decisiones judiciales deben basarse en pruebas objetivas, sin discriminación ni sesgos. Estos marcos jurídicos no surgieron de forma espontánea, sino como producto de largos procesos históricos que buscaron garantizar la protección de la dignidad humana y el bien común. En el caso mexicano, la Constitución de 1917 ha sido el pilar que por más de un siglo ha sostenido el orden jurídico y social del país.

Por ello, permitir que una inteligencia artificial sesgada, entrenada con información incompleta o discriminatoria, influya de forma decisiva en una sentencia penal, representa una regresión. La tecnología no debe desplazar al juicio humano, sino complementarlo bajo criterios éticos claros y con mecanismos de rendición de cuentas. El eje de toda decisión debe seguir siendo la persona, su dignidad y sus derechos.

Si un sistema algorítmico falla, sus consecuencias pueden ser devastadoras. Una condena injusta, basada en una

nerative-ai-the-n ext-productivity-frontier.

predicción errónea, no solo afecta la libertad individual, sino también el proyecto de vida de quien la sufre. Un antecedente penal mal asignado reduce oportunidades laborales y puede tener un impacto psicológico profundo. Además, el uso sostenido de tecnologías injustas puede erosionar la confianza pública en la inteligencia artificial, desincentivando su desarrollo y adopción social.

Por eso, cualquier avance tecnológico debe estar acompañado de una reflexión jurídica y ética. La inteligencia artificial no debe convertirse en un nuevo sesgo estructural ni en una herramienta de discriminación moderna, debe, por el contrario, ser diseñada y utilizada para reforzar los valores de justicia, igualdad y libertad.

Frente a este problema, para conocer cómo es que una IA toma decisiones, la Agencia de los Derechos Fundamentales de la Unión Europea propone una alternativa relevante, permitir que auditores y expertos puedan analizar estos algoritmos bajo acuerdos de confidencialidad. Esta medida permitiría realizar revisiones técnicas sin comprometer los secretos comerciales, pero garantizando la rendición de cuentas.

Esta propuesta adquiere particular relevancia cuando son los propios gobiernos o instituciones de justicia quienes hacen uso de sistemas algorítmicos para tomar decisiones que afectan directamente la vida de las personas. En tales casos, resulta imprescindible asegurar que los mecanismos utilizados estén alineados con los principios de justicia y respeto a los derechos humanos. Solo así se podrá garantizar que ni la tecnología ni los funcionarios públicos actúen con prejuicios o vulneren la dignidad de los individuos.

Esto representa un primer paso relevante, pero debe ir acompañada de un marco regulatorio robusto que defina con claridad cómo debe operar la inteligencia artificial. Si bien una reforma legal puede parecer compleja y lenta, el crecimiento exponencial de esta tecnología exige una respuesta urgente. Es imprescindible establecer una legislación que obligue a todas las empresas que provean sistemas de IA al Estado, en ámbitos como justicia, salud o seguridad, a cumplir con requisitos mínimos de transparencia algorítmica.

Esto no implicaría divulgar secretos industriales, sino informar aspectos fundamentales, qué tipo de bases de datos se utilizan, si se han identificado sesgos y cómo se han corregido, los criterios generales de decisión, las limitaciones del sistema y las precauciones éticas adoptadas. Esta

información podría entregarse de forma periódica, por ejemplo, cada trimestre, a una institución técnica especializada que, bajo un acuerdo de confidencialidad, revise, valide y sugiera mejoras. Este proceso permitiría mantener estándares éticos y técnicos más sólidos, sin comprometer la propiedad intelectual de las empresas.

Aun con estas medidas, debe subrayarse que los sistemas de inteligencia artificial no deben ser nunca el único criterio para tomar decisiones judiciales. Deben integrarse como herramientas auxiliares, sujetas al juicio humano, al análisis crítico y a otras fuentes de evidencia.

En conclusión, la inteligencia artificial ha demostrado ser una herramienta valiosa para mejorar procesos y aumentar la eficiencia, pero cuando opera con sesgos, no solo reproduce desigualdades históricas, sino que las amplifica. Por ello, se requiere una ética algorítmica activa, orientada a garantizar que su uso sea objetivo, justo y centrado en la dignidad humana, promoviendo siempre el bienestar de quienes se ven afectados por sus decisiones.

En el caso de México, el *Global Index on Responsible AI 2024* lo posiciona en el lugar 64 de 138 países, con una puntuación general de 15.77 puntos sobre 100, lo que lo ubica en el cuarto cuartil del ranking global. Este resultado refleja debilidades significativas en la creación de marcos regulatorios, acciones gubernamentales concretas y capacidades institucionales. La posición de México pone en evidencia la urgencia de fortalecer su ecosistema de gobernanza algorítmica para evitar que la IA profundice brechas estructurales y, en cambio, se convierta en una herramienta ética al servicio del bien común.

Referencias

- Angwin, Julia, Jeff Larson, Surya Mattu y Lauren Kirchner. "Machine Bias". ProPublica, 23 de abril de 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
- "Breve historia visual de la inteligencia artificial". National Geographic España, 2025. https://www.nationalgeographic.com.es/ciencia/breve-historia-visual-inteligencia-artificial_1441 9.
- Chui, Michael, Eric Hazan, Roger Roberts et al. "The economic potential of generative AI: The next productivity frontier". McKinsey & Company, 13 de junio de 2023.

https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of generative-ai-the-next-productivity-frontier.

- CÁMARA DE DIPUTADOS DEL H. CONGRESO DE LA UNIÓN. “CONSTITUCIÓN POLÍTICA DE LOS ESTADOS UNIDOS MEXICANOS. https://www.diputados.gob.mx/LeyesBiblio/pdf/CPEUM.pdf
- Dressel, Julia y Hany Farid. "The accuracy, fairness, and limits of predicting recidivism". *Science Advances* 4, n.º 1 (2018): eaao5580. https://doi.org/10.1126/sciadv.aao5580.
- Escobar, Dylan. "Conoce la historia del significado de la palabra robot: tiene un pasado oscuro". infobae, 14 de agosto de 2024. https://www.infobae.com/tecno/2024/08/14/conoce-la-historia-del-significado-de-la-palabra-robo t-tiene-un-pasado-oscuro/.
- European Union Agency for Fundamental Rights. *Bias in algorithms - Artificial intelligence and discrimination*. 2022.
- European Union Agency for Fundamental Rights. *Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights*. 2019. 24. https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making
- European Union Agency for Fundamental Rights. *#BigData: Discrimination in data-supported decision making*. 2018. 24. https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making 11
- Ferrara, Emilio. "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies". *Sci* 6, n.º 1 (2023): 3. https://doi.org/10.3390/sci6010003.
- Historia y Vida. "Alan Turing, el padre del ordenador: todas sus aportaciones". La Vanguardia, 21 de marzo de 2024. https://www.lavanguardia.com/historiayvida/historia-contemporanea/20180627/47312986353/qu e-aporto-ciencia-alan-turing.html.
- Kirkpatrick, Keith. "It's not the algorithm, it's the data". *Communications of the ACM* 60, n.º 2 (2017): 21–23. https://doi.org/10.1145/3022181.
- "La Declaración Universal de los Derechos Humanos". ONU. Consultado el 16 de mayo de 2025. https://www.un.org/es/about-us/universal-declaration-of-human-rights.
- "La IA permite a los trabajadores ahorrar un promedio de una hora, según encuesta". Forbes Centroamérica, 24 de octubre de 2024. https://forbescentroamerica.com/2024/10/18/la-ia-permite-a-los-trabajadores-a

horrar-un-promedi o-de-una-hora-segun-encuesta.
- Nielsen, Jakob. "AI Improves Employee Productivity by 66%". Nielsen Norman Group, 16 de julio de 2023. https://www.nngroup.com/articles/ai-tools-productivity-gains/.
- "QS World Future Skills Index | QS". QS, 2025. https://www.qs.com/insights/world-future-skills/.
- "Qué es la misteriosa “caja negra” de la inteligencia artificial que desconcierta a los expertos (y por qué aún no entendemos cómo aprenden las máquinas) - BBC News Mundo". BBC News Mundo, 2023. https://www.bbc.com/mundo/noticias-65331262.
- Shelley, Mary Wollstonecraft y Cristian Valle. *Frankenstein: El Moderno Prometeo*. Independently published, 2021.

IA sin reglas: ¿avance o amenaza?

Autores: Daniel Hernández Gutiérrez y Diego Barragán Castillo

La inteligencia artificial se ha convertido en una de las herramientas más poderosas de nuestros tiempos. Está presente en múltiples aspectos de la vida cotidiana, desde todas las recomendaciones que vemos en redes sociales mediante algoritmos especializados en mostrar cosas de interés para las diferentes personas, hasta sistemas complejos de diagnósticos médicos, reconocimiento fácil, vigilancia, etc. “La inteligencia artificial (IA) es una tecnología que permite a las computadoras y máquinas simular el aprendizaje humano, la comprensión, la resolución de problemas, la toma de decisiones, la creatividad y la autonomía” (Stryker & Kavlakoglu, 2024). A medida que esta tecnología avanza, también surgen diferentes cuestionamientos sobre los riesgos que puede implicar para los derechos humanos, como la privacidad, la libertad de expresión, la no discriminación, y algo muy importante la no distribución de datos personales. En el *Global Index on Responsible AI 2024* se evidencia que “se encontró que pocos países cuentan con mecanismos para proteger los derechos humanos en riesgo por la IA”. Así, mientras algunos celebran el potencial de la inteligencia artificial en áreas como la salud, la educación, la innovación, la seguridad o el medio ambiente, otros advierten que si no se regula adecuadamente, podría profundizar aún más en las desigualdades ya existentes en la sociedad y esto provocaría una mayor brecha entre quienes tienen acceso a estas nuevas tecnologías y quienes no. Las decisiones automatizadas podrían seguir expandiendo los sesgos ya existentes, los empleos se podrían transformar o desaparecer sin tener respaldo juntos para los empleados, y el poder se concentraría únicamente en unas pocas empresas tecnológicas con demasiado poder a nivel mundial, lo que podría limitar la equidad y la democracia.

En la actualidad, el avance acelerado de inteligencia artificial ha generado distintas posturas, por un lado están quienes defienden la regulación estricta para poder garantizar la protección y el respeto de los derechos humanos, y por el otro lado están quienes temen que dichas restricciones frenen el progreso humano y la innovación tecnológica y limiten los beneficios que esta tecnología

podría ofrecer a miles de personas, así surge un problema moral; por lo tanto, ¿Deben los gobiernos priorizar la regulación estricta de la IA para proteger los derechos humanos, incluso si esto podría ralentizar la innovación y el desarrollo tecnológico?

Este problema nos conduce también a un dilema ético que va más allá de una simple elección entre regulación y libertad tecnológica. Implica reflexionar qué valores deben guiar el desarrollo de la inteligencia artificial en nuestra sociedad. Estos valores pueden incluir el respeto a la dignidad humana, la equidad, la transparencia y la no discriminación. ¿Es más importante priorizar la protección de los derechos humanos a través de una regulación estricta de la IA, o fomentar la innovación tecnológica sin restricciones que podría beneficiar económicamente a la sociedad? Este dilema enfrenta dos principios fundamentales: la justicia y la equidad frente a la eficiencia y el progreso. Resolverlo implica tomar una postura sobre qué tipo de futuro queremos construir, uno guiado por la responsabilidad y la ética, o uno impulsado por la innovación, el cambio y el desarrollo.

La inteligencia artificial ha dejado de ser una tecnología futurista ya que hoy en día es una herramienta muy influyente en múltiples ámbitos de la vida cotidiana. En este contexto, los gobiernos tienen la responsabilidad ética y política de implementar marcos que sean regulatorios y estrictos que garanticen la protección de los derechos humanos ante cualquier uso que se le de. En el *Global Index on Responsible AI 2024* se señala que "las tecnologías no son neutrales: reflejan las intenciones, valores y prejuicios de quienes desarrollan y de quienes la regulan" lo que indica la urgencia de una regulación que impida estos actos por parte de las personas que están a cargo y que dan las órdenes, para así poder reducir las desigualdades estructurales. Esta regulación es particularmente necesaria para evitar la toma de decisiones discriminatorias que pueden perjudicar a grupos históricamente marginados.

Desde la perspectiva de la teoría de la justicia, propuesta por John Rawls, el principio de equidad exige que las instituciones sociales, incluidas las tecnologías, deben de estar diseñadas de modo que beneficien a los menos favorecidos. Según Rawls, "las desigualdades sociales y económicas deben ser arregladas de tal manera que beneficien a los más desfavorecidos, es decir, que mejoren su situación en comparación con el status quo" (Rawls, 1971, p. 102). Este principio se fundamenta en la idea de que las políticas y estructuras deben promover el bienestar de aquellos que se encuentran

en una posición más vulnerable, de modo que las diferencias económicas y sociales no se profundicen. Si se permite que las empresas tecnológicas operen sin ningún tipo de límite, existe el riesgo de que la inteligencia artificial reproduzca e incluso amplifique las brechas e injusticias sociales existentes, en lugar de contribuir a reducirlas. En este sentido, una regulación justa no sería un obstáculo para el desarrollo, sino una condición necesaria para que dicho desarrollo sea inclusivo, legítimo y éticamente sostenible. Esta visión se alinea con la teoría de la justicia de John Rawls, quien sostiene que las políticas deben diseñarse de manera que beneficien especialmente a los más desfavorecidos. Aplicado al contexto tecnológico, esto implica que la regulación justa debe asegurar que los avances en inteligencia artificial mejoren las condiciones de quienes están en desventaja, evitando que las innovaciones se conviertan en herramientas de exclusión o inequidad, y contribuyendo así al desarrollo inclusivo que se busca.

Un ejemplo muy claro de la necesidad de esta regulación de la inteligencia artificial es el caso del sistema “COMPAS” (Correctional Offender Management Profiling for Alternative Sanctions) en Estados Unidos, un algoritmo utilizado para predecir si las personas que ya han sido arrestadas pueden volver a cometer un delito. Este sistema fue denunciado por tener sesgos raciales, ya que si se comparaba a una persona afroamericana con una persona blanca con el mismo historial, la misma edad, el mismo género y el mismo “futuro criminal”, el acusado afroamericano tenía un 45% más de posibilidades de obtener un puntaje de riesgo que un acusado blanco (Maybin, 2016). Claramente se evidencia como la ausencia de regulación permitieron una injusticia a partir de los sesgos algorítmicos y de perpetuación de violencia estructural, demostrando así los peligros de delegar decisiones críticas a sistemas sin un control adecuado. Si este fuera el caso de México las personas con más propensas a sufrir las consecuencias de estos sesgos serían aquellas que pertenecen a comunidades indígenas, las personas en situación de pobreza, reos y quienes viven en zonas marginadas, serían clasificadas como personas riesgosas para la sociedad sin que sea necesariamente el caso, ya que los datos que hay apuntan mucho más a este sector de la población. Dichos datos vienen de la recopilación existente de las personas riesgosas, siendo en su mayoría registros de personas que pertenecen a dichas comunidades. “De acuerdo con el estatus jurídico de las personas privadas de la libertad en los centros penitenciarios federales y estatales*, 86 302 (40.9%) se encontraban Sin sentencia / Medida cautelar de internamiento preventivo, 30 388 (14.4%) con sentencia no definitiv” (INEGI, 2021). El hecho de

que los registros tengan en su mayoría datos de personas que pertenecen a estos grupos o sectores sociales significa que no se han capturado correctamente los datos de los demás sectores involucrados y hay un desbalance en los datos que sesga todo el sistema.

La regulación estricta de la IA no solo es necesaria para corregir sesgos históricos, sino para prevenir nuevos riesgos sistémicos que podrían suceder en caso de una mala supervisión. El Reglamento de IA de la Unión Europea, vigente desde febrero de 2025, prohíbe prácticas como la manipulación subliminal de los datos o durante el entrenamiento de la IA, la explotación de vulnerabilidades socioeconómicas y el reconocimiento facial no autorizado (*Regulación De La IA: Prohibiciones a Partir De 2025*, 2025). Estas medidas son consideradas justas y adecuadas, que responden a casos documentados de discriminación algorítmica que han sucedido recientemente y que levantan controversía sobre el manejo y regulaciones que debería de tener la IA, como sistemas de "puntuación social" que profundizan desigualdades.

La abogada Vanesa Alarcón, asesora de empresas y startups de tecnología para la protección de datos (AGM Abogados), advierte que sin marcos legales "garantistas", las empresas priorizan la eficiencia de los sistemas basados en IA sobre los derechos fundamentales de las personas, siendo un ejemplo de estos la seguridad, privacidad y autonomía, especialmente en grupos marginados, minorías o históricamente discriminados (Quintana, 2025).

Un ejemplo crítico es la prohibición europea de inferir emociones en entornos laborales y educativos. ya que viola la privacidad y/o emocional de las personas, una práctica común en herramientas de análisis de productividad que estigmatizan a empleados y alumnos, con ansiedad o estrés durante su jornada laboral o su horario de clases (Quintana, 2025). Esto refleja el principio de Rawls, que se centra en la justicia social e igualdad de oportunidades para todos, para así proteger a los más vulnerables, si la IA replica los sesgos humanos que pueden existir dentro de quién las programa o entrena, su regulación debe ser tan rigurosa como las leyes antidiscriminatorias que ya existen y así garantizar un trato igualitario y equitativo de todas las personas que se vean afectadas por su uso o implementación. Además, el reglamento exige transparencia en IA generativa a partir de que entre en vigor en agosto de 2025, lo que permitirá auditar sistemas como Chat GPT para evitar desinformación masiva o a las empresas desarrolladoras de dichos sistemas, en este caso OpenAI.

La postura de "legal by design" que propone la Unión Europea alinea la innovación con la ética desde su concepción, para así mantener a la IA neutral y sin sesgos en la medida de lo posible, no como un parche posterior a una regulación o a un evento que los orille a generar dicho parche. Esto es vital en sectores como la biomedicina, donde algoritmos como AlphaFold de Google, aunque beneficioso el algoritmo, podrían usarse para diferentes propósitos muy diferentes a los iniciales u originales, por ejemplo si AlphaFold sufre un mal uso de su sistema de IA, pudiera darse el caso donde se creen amenazas biológicas, donde si existen controles y regulaciones esto no sucedería. La regulación no frena ni perjudica el progreso o su desarrollo, sino que canaliza sus aplicaciones hacia el bien común de las personas (Quintana, 2025).

En un mundo donde la inteligencia artificial representa una de las innovaciones más revolucionarias y potentes del progreso económico y científico, imponer regulaciones estrictas desde los gobiernos podría frenar significativamente la innovación, limitar la competitividad global y retrasar soluciones tecnológicas esenciales como el área de salud en la que hay demasiadas enfermedades que aún no tiene cura y que se podrían encontrar con la inteligencia artificial, otra área es la educación, las nuevas generaciones tienen y van a tener que dominar por completo la IA, y el área del cambio climático que es urgente de frenar. La regulación excesiva, especialmente en etapas del desarrollo de la IA, puede limitar la creatividad de los investigadores y emprendedores y obstaculizar la implementación de sistemas que podrían traer beneficios masivos e increíbles para toda la humanidad.

Desde una perspectiva utilitarista, lo correcto es aquello que genera el mayor bienestar para el mayor número de personas. En este sentido, permitir que la innovación tecnológica avance con agilidad puede resultar más beneficioso a largo plazo para toda la sociedad. John Stuart Mill sostiene que "las acciones son correctas en la medida en que tienden a promover la felicidad, equivocadas en cuanto tienden a producir lo contrario de la felicidad" (*Utilitarianism*, 2009). De esta manera, frenar el avance de tecnologías con un gran potencial de cambiar el mundo y la forma en que vivimos solo por temor a posibles riesgos podría traer una pérdida significativa de oportunidades para mejorar la vida de millones de personas. Los avances de la IA han contribuido al diagnóstico temprano de enfermedades, acceso a información, resolución de problemas, entre muchas otras cosas. El filósofo Peter Singer señala que "el principio utilitarista exige que tomemos en cuenta

todos los intereses afectados por nuestras acciones, y actuemos para lograr el mejor equilibrio posible entre ellos" (*Practical Ethics*, 2011). Bajo esta referencia, el desarrollo de la inteligencia artificial puede considerarse éticamente correcto si su uso es responsable y produce muchos más beneficios que perjuicios para la sociedad y el mundo.

Un gran ejemplo para evidenciar la innovación sin regulaciones limitantes es el desarrollo de AlphaFold, un sistema desarrollado por DeepMind que emplea redes neuronales para predecir con exactitud la estructura tridimensional de casi todas las proteínas conocidas a partir de sus secuencias de aminoácidos (DataScientist, 2024). Además, AlphaFold predice el comportamiento de moléculas que se utilizan habitualmente en fármacos, como son los ligandos y los anticuerpos. Estas moléculas se unen a determinadas proteínas y cambian el modo en que estas afectan a la salud humana o al desarrollo de enfermedades. Esta herramienta revolucionó por completo la biología molecular y fue liberada gratuitamente para la comunidad científica y ayuda a los científicos a formular nuevas hipótesis que después ponen a prueba en sus laboratorios (Google DeepMind AlphaFold team, 2024). Si este tipo de avances hubiera estado restringido por normativas excesivamente estrictas, posiblemente no habría tenido el impacto que tiene hoy en día en la investigación médica, farmacéutica y biotecnológica, demostrando cómo la libertad en la innovación puede generar beneficios globales.

La regulación excesiva podría asfixiar el potencial de la IA para resolver crisis que alcancen la magnitud global. El Programa Ejecutivo en IA de ISDI señala que el ritmo de avance tecnológico que existe, supera la capacidad de los legisladores para comprenderlo, lo que lleva a normas desactualizadas que obstaculizan proyectos legítimos y que están dentro de lo ético en la mayoría de los casos (ISDI, 2024). Por ejemplo, la prohibición europea de usar IA para predecir delitos ignora su utilidad en prevenir diferentes ataques que se pudieran llegar a presentar, como por ejemplo ataques terroristas mediante un análisis de patrones de las redes sociales, siempre que se combine con una buena supervisión humana para así evitar riesgos y sesgos, lo que contribuiría a salvar muchas vidas.

El modelo de autorregulación empresarial propuesto por Marco Argenti (Goldman Sachs) muestra que la IA ya está generando "empleados digitales" que aumentan la productividad en equipos híbridos, estos empleados son el futuro de la organización de las empresas y las dinámicas que tendrán, al ser digitales pueden ayudar a maximizar la

productividad, eficiencia y ganancias de las empresas (Digital Robots, 2025). Si la Unión Europea hubiera impuesto sus actuales restricciones en 2020, herramientas como GitHub o Copilot, que hoy ayudan a millones de programadores, estudiantes, trabajadores, maestros, entre otros, no existirían. La filosofía utilitarista de Mill justifica este enfoque que buscan las empresas con los sistemas de IA generativa, un muy buen ejemplo sería el beneficio colectivo de acelerar diagnósticos médicos con IA, como lo fue durante la pandemia de COVID-19, donde se utilizaron para distinguir síntomas, ya que eran muy similares a otros malestares de manera inicial, y así poder atender de manera correcta a los pacientes, superando los riesgos teóricos de privacidad médica, ya que no se compartían dichos datos y se mantenían dentro de la privacidad del doctor, hospital y paciente.

Además, la regulación localizada que plantean las empresas emergentes o ya establecidas, adaptando IA a contextos culturales, sociales o geopolíticos específicos, es más efectiva que normas globales rígidas que puedan limitar el potencial, desarrollo y alcance de la IA en ciertas regiones. Por ejemplo, en agricultura, algoritmos que no están regulados, permiten a pequeños productores africanos optimizar sus cosechas usando datos hiperlocales, algo que burocracias centralizadas no podrían gestionar ni aprobar y esto afectaría al sector agrícola en este caso. Frenar esto con legislaciones prematuras y mal organizadas, perpetuaría la dependencia tecnológica que existe alrededor del mundo y de la cuál muchas personas se ven beneficiadas, como por ejemplo en el Sur Global (Digital Robots, 2025).

Para concluir nuestro dilema ético principal: ¿Deben los gobiernos priorizar la regulación estricta de la IA para proteger los derechos humanos, incluso si esto podría ralentizar la innovación y el desarrollo tecnológico? Debemos revisar ambas posturas de dicho dilema. Tomando en cuenta la parte en la que se toman a favor las regulaciones estrictas donde podemos ver situaciones en las que las personas se han visto afectadas por la falta de respeto hacia algunos de sus derechos básicos, como lo son la privacidad, y seguridad, violando así la integridad de las personas como lo hemos visto al momento de hacer reconocimientos faciales no consensuados o la discriminación racial que existe en otros sistemas que se han implementado para verificar si una persona podría o no volver a cometer un crimen después de haber estado en prisión. Por ello las regulaciones estrictas evitarían dichos acontecimientos, para así proteger la integridad de las personas. Estas regulaciones buscan disminuir las discriminaciones socioeconómicas que se

pueden presentar durante las implementaciones de los sistemas basados en IA generativa. Por otro lado, tomamos en cuenta la postura donde las regulaciones no deberían de ser tan estrictas y ser un poco más liberales en cuanto a los sistemas basados en IA. Estas regulaciones podrían ayudar a contribuir al desarrollo y alcance de dichos sistemas, los cuales se han visto benéficos en muchas otras situaciones donde se han implementado.

Por último, nuestra postura recae en que las regulaciones deberían ser localizadas, dando así la decisión de regulación a cada país manteniendo la integridad de su cultura, sociedad y necesidades geopolíticas que necesite cada uno. Permitiendo así un mayor alcance, progreso y desarrollo de la IA en diferentes ámbitos cotidianos, económicos, empresariales y/o sociales, para así tener un beneficio enfocado en el bien común de las personas que la utilicen. Dando así una solución al dilema ético donde los gobiernos prioricen la efectividad y optimización de dichos sistemas pensando a futuro, la mayoría de los progresos tecnológicos que han sido implementados en el mundo han pasado por el mismo dilema y el gobierno los ha permitido a lo largo de la historia, ¿por qué este caso debería de ser diferente? No debería, entonces las posturas gubernamentales necesitan recaer en el progreso sacrificando de manera mínima la integridad de sus ciudadanos por el beneficio de ellos a futuro.

Las consecuencias de está postura es que si en algún caso de que algún gobierno o empresa no respete estás regulaciones localizadas serían la creación de un organismo global que se encargue de revisar dichas regulaciones localizadas para así prevenir sesgos que se pudieran presentar, en caso de que algún país decida no acoplarse a dichos parámetros podría ser auditado para la revisión y corrección de sus regulaciones. Adicionalmente a esto, las innovaciones y progresos que se vean en diferentes países deberán ser compartidos e implementados donde las regulaciones lo permitan para así mantener una justicia social de manera mundial y una igualdad de oportunidades que sean en beneficio de todos.

Referencias

- Adams, R., Adeleke, F., Florido, A., de Magalhães Santos, L. G., Grossman, N., Junck, L., & Stone, K. (2024). Global Index on Responsible AI 2024 (1st Edition). Global Center on AI Governance. https://coral-trista-52.tiiny.site/
- *Article 5: Prohibited AI Practices | EU Artificial Intelligence Act*. (2025, 02 2). EU AI Act.

https://artificialintelligenceact.eu/article/5/

- Data Scientist. (2024). AlphaFold: Todo lo que necesitas saber. https://datascientest.com/es/alphafold-todo-lo-que-necesitas-saber
- Google DeepMind AlphaFold team. (2024). AlphaFold 3 predice la estructura y las interacciones de todas las moléculas de la vida. Google. https://blog.google/intl/es-es/noticias-compania/iniciativas/alphafold-3-predice-la-estructura-y-las-interacciones-de-todas-las-moleculas-de-la-vida/#:~:text=AlphaFold%203%20predice%20el%20comportamiento,o%20al%20desarrollo%20de%20enfermedades.
- INEGI. (2022). Censo Nacional de Sistema Penitenciario Federal y Estatales 2021. https://www.inegi.org.mx/contenidos/programas/cnspef/2021/doc/cnsipef_2021_resulta dos.pdf
- *La Revolución de la Inteligencia Artificial en 2025: Un Futuro Transformador — Digital Robots*. (2025). Digital Robots.https://www.digital-robots.com/noticias/la-revolucion-de-la-inteligencia-artifi cial-en-2025-un-futuro-transformador
- Maybin, S. (2016). ¿Cómo en Estados Unidos las matemáticas te pueden meter en prisión? BBC News. https://www.bbc.com/mundo/noticias-37679463
- RAWLS, J. (1971). *A Theory of Justice: Original Edition*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9z6v
- *Regulación de la IA: marco legal, desafíos y perspectivas futuras*. (2024). ISDI. https://www.isdi.education/es/blog/regulacion-de-la-ia-2
- *Regulación de la IA: Prohibiciones a partir de 2025*. (2025). Consultoría LOPD. https://gestionalopd.es/regulacion-de-la-ia-prohibiciones-a-partir-de-2025/
- Stuart, J. (2009). *Utilitarianism*. The Floating Press. https://www.utilitarianism.com/jsmill-utilitarianism.pdf
- Singer, P. (2011). *Practical Ethics Third Edition.* Cambridge. https://oldmis.kp.ac.rw/admin/admin_panel/kp_lms/files/digital/Core%20Books/Philos ophy%20&%20psychology/Practical-ethics.pdf
- Stryker, C. & Kavlakoglu, E. (2024). ¿Qué es la inteligencia artificial (IA)? IBM. https://www.ibm.com/mx-es/think/topics/artificial-intelligence
- Quintana, G. (2025). *Europa regula la IA: «En 2025 se exigirá más transparencia en el uso de IA»*. Paréntesis.Media. https://www.parentesis.media/europa-regula-la-ia-impacto-global-y-desafios-para-las empresas/

Deberes y Dilemas: La Gobernanza Ética de la IA y los Derechos Humanos

Autores: Jessica Ángel Galván e Isidoro Salvador Rodriguez Valderrama

La Inteligencia Artificial (IA) ha sido definida de diversas maneras y representa uno de los desarrollos tecnológicos más significativos de nuestra era. Definida por Russell y Norvig (2021) que su objetivo y propósito principal es desarrollar sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el razonamiento, el aprendizaje y la toma de decisiones. Sin embargo, esta definición técnica no captura completamente su impacto transformador en la sociedad contemporánea. La IA ha evolucionado de ser una herramienta de asistencia a convertirse en un actor determinante en decisiones que afectan derechos fundamentales, desde la selección de personal hasta la administración de justicia, configurando así uno de los dilemas éticos más apremiantes en la actualidad.

Con el paso del tiempo, ha comenzado a convertirse en una parte importante de nuestra cultura, y ha adquirido un papel fundamental en nuestras sociedades, impulsada y fomentada por gobiernos, empresas y líderes tecnológicos como una vía eficiente para cumplir objetivos. A medida que esta tecnología se expande, también se vuelve más relevante en nuestra vida diaria y en el diseño de políticas públicas. La eficiencia de esta tecnología nos ha permitido optimizar recursos, mejorar procesos y abrir nuevas posibilidades en áreas como la salud, la educación, la industria y la comunicación. Convirtiéndose en una herramienta que nos ha permitido evolucionar, romper barreras y generar avances maravillosos en tecnología e innovación, incluso ayudándonos a imaginar futuros más sostenibles e innovadores. En este sentido, la Inteligencia Artificial ha sido una aliada en la creación de espacios para el desarrollo humano, permitiendo avances en tantas áreas que hace apenas unas décadas nos hubieran parecido inimaginables.

Sin embargo, y a pesar de sus beneficios, su creciente presencia y expansión también ha generado importantes cuestionamientos éticos y morales. Aunque su propósito inicial es el asistir al ser humano, su implementación ha dado

lugar a dilemas sobre el uso correcto de esta tecnología ya que este no siempre ha estado alineado con principios de justicia, equidad o protección de los derechos humanos. ¿Qué pasa cuando una herramienta tan poderosa se despliega sin suficientes límites o regulaciones? Aún más cuando esta no es capaz de ser responsable por sí misma, sino dependiente y responsabilidad de quienes la diseñan, la usan o se benefician de ella.

Un claro ejemplo de esta tensión lo presenta el Índice Global de IA Responsable 2024, revela un problemática preocupante, que expone cómo, a pesar de que el 39% de los países evaluados cuentan con estrategias nacionales de IA, la mayoría carecen de mecanismos vinculantes y no incluyen principios claros de Inteligencia Artificial responsable, lo que implica que en muchos casos, las políticas existentes no garantizan un uso ético ni respetuoso de los derechos humanos, como lo señala el reporte de ILDA “la gobernanza de la IA sigue siendo más una idea que una práctica, fallando en asegurar la implementación responsable de IA” (ILDA, 2024). Este tipo de implementación incompleta deja fuera principios fundamentales como el respeto a los derechos humanos, la ética en cada etapa del ciclo de vida tecnológico y la atención a las necesidades sociales de las comunidades a las que la Inteligencia Artificial pretende servir. Esta falta de compromiso efectivo pone en evidencia que no basta con tener estrategias bien redactadas; es necesario que estas se traduzcan en acciones concretas que protejan a las personas.

Las consecuencias de una IA no regulada o gestionada adecuadamente son preocupantes y se manifiestan en violaciones a derechos fundamentales, como la invasión a la privacidad, la pérdida de empleos y la discriminación algorítmica. Esta realidad subraya el riesgo inherente a una aplicación irresponsable de la tecnología.

Esta introducción nos lleva a plantear un problema moral:

¿Cuál es la responsabilidad moral de los gobiernos al desarrollar estrategia de IA sin mecanismo vinculantes, considerando los riesgos potenciales para un uso irresponsable de la tecnología y la afectación de los derechos humanos?

Para abordar la compleja cuestión de la responsabilidad o el deber moral de los gobiernos al desarrollar estrategias en torno a la Inteligencia Artificial sin mecanismos vinculantes, es necesario examinar no sólo el impacto práctico de estas decisiones, sino también los principios éticos que deben guiar el uso

de una tecnología tan poderosa. Desde una perspectiva deontológica, que se centra en el cumplimiento de deberes y obligaciones morales independientemente de las consecuencias, el análisis se centra en los deberes y obligaciones morales inherentes a la función gubernamental. Especialmente cuando se trata de una tecnología como la IA con una capacidad de influir profundamente en la vida de las personas el Estado tiene el deber de proteger a sus ciudadanos.

La deontología, especialmente en la tradición kantiana, subraya la importancia de tratar a las personas como fines en sí mismas y no como simples medios. La protección de los derechos humanos (privacidad, igualdad, no discriminación) es un imperativo categórico en la deontología. Cuando la IA, sin regulación, pone en riesgo estos derechos, se está violando un deber moral fundamental del gobierno de asegurar la dignidad humana. Tomando en cuenta también, que una regulación deontológica de esta tecnología buscaría establecer principios que puedan aplicarse universalmente, garantizando que el uso de esta misma sea justo y equitativo para todos, en lugar de depender de resultados inciertos o contextos específicos.

La Inteligencia Artificial, lejos de ser una herramienta neutral, está moldeada por decisiones humanas, que reflejan valores, intereses y prioridades. Debido a su fuerte influencia en la vida de las personas, exige un marco ético que guíe su desarrollo y aplicación. En este sentido, el simple desarrollo de estrategias no garantiza una implementación responsable ni justa de esta misma. Las políticas no vinculantes, es decir, aquellas que carecen de obligaciones legales claras y mecanismos de seguimiento, corren el riesgo de convertirse en meras declaraciones de buenas intenciones sin impacto real. Esto es particularmente grave cuando se trata de tecnología, ya que la ausencia de un marco regulatorio sólido permite que la IA afecte derechos fundamentales como la privacidad, la igualdad de oportunidades, el acceso a servicios e incluso la libertad de expresión, lo cuál plantea una seria interrogante sobre el cumplimiento del deber de protección de los ciudadanos por parte del Estado.

Según el artículo 6. del Código de Ética de las personas Servidoras Públicas del Gobierno Federal del Gobierno de México, “la ética pública se rige por la aplicación de los Principio Constitucionales de Legalidad, Honradez, Lealtad, Imparcialidad y Eficiencia” por lo que, es el deber de los gobiernos proteger a sus ciudadanos, lo que implica anticiparse a los posibles riesgos y establecer límites cuando estos sean necesarios. Dejar a las empresas tecnológicas o a los intereses privados la

responsabilidad de autorregularse en materia de IA ha demostrado ser insuficiente. Múltiples casos donde la autorregulación ha fallado como la recopilación masiva de datos sin consentimiento o el uso de IA en vigilancia sin supervisión adecuada, han demostrado cómo los sistemas algorítmicos pueden perpetuar sesgos, excluir a grupos vulnerables o ser utilizados para vigilar y controlar a la población sin garantías adecuadas.

Como el caso de la IA "COMPAS" de 2015 realizada por Northpointe, un sistema que determinaba si un recluso era capaz de repetir un crimen o no, mostrando una afectación principal hacia los grupos afroamericanos, mostró que un 44.9% de los presos afroamericanos si iban a rescindir en la actividad ilícita, cuando en realidad, no rescindieron. En este contexto, los gobiernos tienen una responsabilidad activa, no solo de fomentar el desarrollo tecnológico, sino de hacerlo bajo marcos éticos que aseguren el bienestar común.

La ausencia de mecanismos vinculantes puede interpretarse, más allá de lo inadecuado, como una omisión moral. Al no establecer regulaciones claras y exigibles, los gobiernos permiten (o al menos no demuestran un alto a ello) que la IA sea utilizada de manera irresponsable. Esto puede derivar en consecuencias graves, como la normalización de prácticas discriminatorias, la automatización de decisiones injustas o la erosión progresiva de derechos fundamentales. En otras palabras, no se trata únicamente de una falla técnica o legal, sino de una falta de compromiso ético con la protección de las personas.

Por otro lado, también hay quienes sostienen que las estrategias flexibles permiten una mayor adaptabilidad y fomentan la innovación. Según esta perspectiva, imponer restricciones demasiado rígidas desde el inicio podría frenar el avance tecnológico y limitar su potencial para resolver grandes desafíos sociales. Sin embargo, esta visión olvida que la innovación sin responsabilidad puede resultar más perjudicial que beneficiosa. Las empresas priorizan la innovación y el beneficio económico, lo cuál no siempre se alinea con los principios de equidad, transparencia o la protección de datos. Desde una perspectiva deontológica, la innovación no es un imperativo categórico que justifique la violación de los derechos humanos. El desarrollo ético de la tecnología no busca limitar su potencial, sino orientarlo hacia fines justos y sostenibles. La innovación no debe estar reñida con la protección de derechos, sino ser un medio para promoverlos.

En este contexto, la moralidad de las decisiones gubernamentales no puede

evaluarse sólo en función de sus intenciones declaradas, sino de sus efectos concretos. Si las estrategias de IA no se traducen en marcos legales exigibles, acompañados de mecanismos de rendición de cuentas, participación ciudadana y evaluación constante, entonces los gobiernos están fallando en su deber ético de garantizar el uso responsable de esta tecnología. Por último, es importante considerar que la gobernanza ética de la IA no es una tarea individual ni exclusiva de los gobiernos, pero ellos sí tienen la capacidad (y la obligación como supuestos gobiernos correctos) de establecer las condiciones para que los demás actores (empresas, desarrolladores, usuarios, etc...) actúen también con responsabilidad. No es moralmente aceptable que los gobiernos se limiten a observar desde la distancia, confiando en que los riesgos se gestionan solos. Su ausencia de acción, o su acción insuficiente, los hace cómplices de los daños que puedan derivarse del uso irresponsable de la IA. Y a partir de esto se deriva un dilema ético que pone sobre la balanza dos valores fundamentales, la innovación y el respeto a los derechos humanos.

Una vez visualizada y analizada la pregunta planteada y clave de este ensayo, se configura un dilema central entre fomentar la innovación tecnológica y garantizar el respeto y la protección de los derechos humanos. Ambos valores son esenciales para el progreso de las sociedades modernas, pero a menudo entran en tensión cuando se trata del desarrollo y uso de tecnologías tan potentes como la inteligencia artificial.

Por un lado, la innovación tecnológica impulsada por la IA se presenta como un motor crucial para el desarrollo económico y la mejora de la calidad de vida, con el potencial de revolucionar sectores desde la salud hasta el transporte. Desde esta perspectiva, estrategias gubernamentales flexibles y no vinculantes son vistas como deseables, pues permitirían a desarrolladores y empresas experimentar y avanzar rápidamente sin las restricciones que podrían frenar la creatividad y la inversión.

No obstante, esta visión, aunque valiosa, subestima el impacto ético de una IA que no es neutral. Construida sobre datos que pueden reflejar desigualdades sociales y decisiones humanas sesgadas, la IA tiene el potencial de reproducir o amplificar estructuras de poder existentes. Por ello, el respeto a derechos humanos fundamentales como la igualdad, la privacidad y la no discriminación no puede ser sacrificado en nombre del progreso y debe establecerse como una base.

La respuesta, desde una ética pública

responsable, debe inclinarse hacia un equilibrio justo entre ambos principios. La innovación no debe ser sacrificada, pero tampoco puede ser vista como un fin en sí mismo si sus medios o consecuencias resultan injustas o dañinas.

En última instancia, este dilema revela que el verdadero desafío no está en elegir entre innovación o derechos humanos, sino en diseñar marcos que los integren de manera
coherente y ética. La regulación no debe ser un obstáculo a la creatividad, sino una guía para asegurarse de que el avance tecnológico se traduzca en un beneficio compartido, sin dejar a nadie atrás. Cuando los gobiernos no asumen este deber, cuando no establecen límites claros ni mecanismos vinculantes, no solo están apostando por una innovación irresponsable, sino que están desprotegiendo activamente a sus ciudadanos, especialmente a los más vulnerables. Y eso, es una omisión ética grave.

¿Es éticamente correcto permitir que la falta de regulaciones vinculantes sobre la Inteligencia Artificial fomente la innovación y el progreso tecnológico, aunque esto suponga un riesgo para la protección de los derechos humanos y la justicia social?

En el contexto actual del acelerado desarrollo tecnológico, la inteligencia artificial se presenta como una de las herramientas más poderosas y transformadoras del siglo XXI. Sus beneficios potenciales son innegables, abarcando desde la mejora en diagnósticos médicos hasta la optimización de sistemas de transporte o la personalización de la educación. Sin embargo, su creciente influencia en la vida pública y privada también ha despertado serias preocupaciones éticas, especialmente cuando su avance se produce en un entorno carente de regulaciones claras. La profundidad de esta interrogante exige un análisis riguroso que, más allá de la mera búsqueda de la innovación, cuestione la legitimidad moral de un progreso que podría socavar principios fundamentales como la justicia, la equidad y la dignidad humana, a los cuales el Estado está moralmente obligado a proteger.

Los sistemas algorítmicos no son entidades neutrales ni objetivas; por el contrario, están construidos sobre datos históricos que reflejan desigualdades existentes y decisiones humanas que incorporan sesgos y valores. Permitir que estos sistemas operen sin marcos éticos ni supervisión estatal efectiva no solo representa una omisión técnica, sino una falla moral. Al dejar la autorregulación en manos de empresas privadas, cuyo interés primario suele ser el beneficio

económico y no la equidad social, se desatienden los principios de justicia, transparencia y rendición de cuentas. La innovación, por sí sola, no justifica vulneraciones a la dignidad humana. El progreso que margina, discrimina o vigila sin garantías adecuadas no es verdaderamente progreso, sino un retroceso envuelto en promesas tecnológicas.

Es cierto que una regulación excesiva o mal diseñada podría sofocar la creatividad y ralentizar el avance científico. Pero esto no debe convertirse en excusa para la inacción. La clave está en encontrar un equilibrio que permita el desarrollo de tecnologías útiles, sin dejar de lado la responsabilidad de velar por su impacto ético y social. No se trata de frenar la innovación, sino de orientarla hacia fines justos y sostenibles. En este sentido, la regulación no es un obstáculo, sino una guía que canaliza el poder transformador de la IA hacia el bienestar colectivo.

El análisis de la gobernanza de la Inteligencia Artificial a través de una lente deontológica revela una verdad fundamental: la responsabilidad moral de los gobiernos trasciende la mera promoción de la innovación. Como hemos explorado, el Estado tiene un deber categórico e ineludible de proteger a sus ciudadanos y salvaguardar sus derechos fundamentales. Este deber, arraigado en principios de justicia, equidad y dignidad humana, es un imperativo moral que no puede ser condicionado por la búsqueda de progreso tecnológico sin límites.

La evidencia presentada por organismos como ILDA y el Índice Global de IA Responsable subraya una preocupante realidad: la prevalencia de estrategias de IA gubernamentales que carecen de mecanismos vinculantes. Esta ausencia no es una simple deficiencia técnica, sino una profunda omisión moral. Al permitir que la autorregulación de la industria prime sobre marcos éticos y legales claros, los gobiernos fallan en su deber de anticipar y mitigar riesgos. Casos como el algoritmo COMPAS son un testimonio contundente de cómo esta pasividad gubernamental conduce a la perpetuación de sesgos, la discriminación y la erosión progresiva de derechos fundamentales, transformando una herramienta de potencial beneficio en una fuente de injusticia.

Como última instancia, el dilema no radica en elegir entre innovación o derechos humanos, sino en la obligación moral de los gobiernos de integrar ambos de manera coherente y ética. La regulación no es un obstáculo a la creatividad, sino una guía indispensable que canaliza el poder transformador de la IA hacia un beneficio compartido y justo. Permitir que la falta de regulaciones

vinculantes fomente una innovación irresponsable no es éticamente aceptable; por el contrario, constituye una desprotección activa de los ciudadanos, especialmente de los más vulnerables. Por lo tanto, la verdadera ética del progreso tecnológico reside en la acción decidida de los gobiernos para establecer límites claros y exigibles, garantizando que el avance de esta tecnología en constante crecimiento y trascendencia sea inclusivo, equitativo y profundamente respetuoso de la dignidad humana. Solo así podremos construir un futuro donde la eficiencia tecnológica se alinee con la justicia social.

Referencias

- The Global Index on Responsible AI. (s. f.). https://www.global-index.ai/ Belver, V. (2024c, agosto 9). La investigación del Índice Global de IA Responsable en América Latina: colaboración y conocimiento en red. ILDA. https://idatosabiertos.org/la-investigacion-del-indice-global-de-ia-responsable-en-a merica-lati na-colaboracion-y-conocimiento-en-red/
- Belver, V. (2024b, junio 25). Se lanzó el Índice Global de IA Responsable. ILDA. https://idatosabiertos.org/se-lanzo-el-indice-global-de-ia-responsable/
- Del Pilar Roa Avella, M., Sanabria-Moyano, J. E., & Dinas-Hurtado, K. (2022). Uso del algoritmo COMPAS en el proceso penal y los riesgos a los derechos humanos. Revista Brasileira de Direito Processual Penal, 8(1). https://doi.org/10.22197/rbdpp.v8i1.615 UNESCO. (2021). Recomendación sobre la ética de la inteligencia artificial. Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. https://unesdoc.unesco.org/ark:/48223/pf0000381137
- Auxiliares, A. y. S. (s. f.). Código de Ética de las personas Servidoras Públicas del Gobierno Federal. gob.mx. https://www.gob.mx/asa/acciones-y-programas/codigo-de-etica-de-los-servidores-p ublicos#:~ :text=La%20%C3%A9tica%20p%C3%BAblica%20se%20rige,y%20reglas%20de%20int egri dad%2C%20que
- ValgrAI. (2023). La ética en la inteligencia artificial: una necesidad urgente. https://valgrai.eu/la-etica-en-la-inteligencia-artificial-una-necesidad-urgente/
- López Velarde, M. A., & Domínguez Pérez, C. E. (2022). Regulación de la inteligencia artificial: Desafíos para los derechos humanos en México. Revista de Derecho Público, (30), 57–80. https://revistas.unam.mx/index.php/derechopub/article/view/80484

Hacia un Futuro Responsable: Ética, Datos e Inteligencia Artificial en Debate

Autor: José Manuel Cuevas Avila

En una era donde las decisiones automatizadas moldean desde el contenido que consumimos hasta nuestras oportunidades laborales o crediticias, la transparencia y responsabilidad ética en la inteligencia artificial (IA) se ha vuelto no solo deseable, sino urgente. Nos enfrentamos a tecnologías que, a pesar de su complejidad técnica y sus beneficios, operan muchas veces como *cajas negras* (es decir, sistemas cuyo funcionamiento interno no es comprensible o accesible para los usuarios). Esta opacidad plantea riesgos éticos serios en contextos donde las decisiones de un algoritmo pueden afectar directamente derechos fundamentales como la igualdad, la privacidad o el debido proceso.

El objetivo de este artículo es analizar, desde una perspectiva ética, por qué resulta imprescindible exigir transparencia y rendición de cuentas en el diseño y aplicación de algoritmos de IA. A lo largo del texto se presentarán ejemplos reales, estudios de caso y evidencia estadística que demuestran cómo la falta de supervisión y explicabilidad en estos sistemas puede perpetuar injusticias, amplificar desigualdades y dificultar la identificación de responsabilidades cuando ocurren errores. La pregunta central que guía esta reflexión es: **¿cuál es la urgencia de exigir transparencia y responsabilidad ética en el desarrollo de algoritmos de inteligencia artificial?**

Uno de los principales desafíos éticos actuales es que muchas decisiones tomadas por IA no pueden ser explicadas adecuadamente. Esto es especialmente grave cuando afectan aspectos sensibles como la justicia, la salud, el empleo o el acceso a servicios. Si una persona es rechazada para un crédito, despedida, arrestada o diagnosticada erróneamente con base en una decisión algorítmica, ¿quién puede explicar qué ocurrió? ¿Quién asume la responsabilidad si el algoritmo cometió un error?

En palabras de la investigadora Cathy O'Neil (2016), los algoritmos mal diseñados pueden convertirse en “armas de destrucción matemática” cuando operan sin rendición de cuentas y afectan a las poblaciones más vulnerables. La falta de mecanismos de revisión ética y la

ausencia de explicabilidad comprometen la justicia de los sistemas automatizados.

Un caso emblemático que ilustra la urgencia de esta discusión es el del sistema COMPAS, utilizado en Estados Unidos para predecir la probabilidad de reincidencia criminal. Este algoritmo influía en decisiones judiciales, como la libertad condicional. Una investigación de *ProPublica* en 2016 reveló que COMPAS presentaba un sesgo racial: clasificaba erróneamente a personas afroamericanas como de alto riesgo casi el doble de veces que a personas blancas, incluso cuando los antecedentes eran similares (Angwin et al., 2016).

Lo más grave no fue solo el sesgo, sino la falta de acceso al funcionamiento del algoritmo. La empresa que lo desarrolló, Northpointe (ahora Equivant), argumentó que su sistema era propietario y no podía revelar su lógica interna. Esto imposibilitó cualquier apelación fundamentada por parte de los acusados, violando principios básicos de transparencia judicial y debido proceso.

Otro ejemplo relevante es el intento de Amazon por automatizar su proceso de reclutamiento mediante un sistema de IA entrenado con currículums de una década. El algoritmo, en lugar de seleccionar al mejor talento, aprendió a penalizar automáticamente a las mujeres en postulaciones para áreas técnicas, porque en los datos históricos predominaban hombres (Hao, 2019). Es decir, el sistema reprodujo los sesgos del pasado sin capacidad crítica ni ajuste ético.

Este caso muestra cómo incluso empresas tecnológicas líderes pueden fallar gravemente cuando no se implementan mecanismos de revisión ética o auditoría algorítmica antes de aplicar estos sistemas a la vida real. Amazon terminó descartando el proyecto, pero nunca hizo pública la lógica completa del sistema ni se sometió a un análisis independiente.

Las redes sociales constituyen uno de los espacios más visibles y cotidianos donde la inteligencia artificial impacta a millones de personas. Plataformas como Instagram, TikTok o Facebook utilizan algoritmos complejos para decidir qué contenido mostrar, cuándo y a quién. Estos sistemas no operan de forma neutral: su objetivo es maximizar el tiempo de permanencia del usuario, lo que genera dinámicas de adicción, exposición selectiva de información e incluso alteraciones en la percepción del cuerpo y la identidad.

En un estudio de *Harvard University's Berkman Klein Center*, se encontró que más del 70% de los adolescentes en Estados Unidos reportan haber cambiado sus hábitos alimenticios, rutinas de ejercicio o productos de consumo debido a recomendaciones de redes sociales

guiadas por algoritmos (Gottfried & Shearer, 2022). Esta influencia se produce muchas veces sin que el usuario comprenda que está interactuando con un sistema diseñado específicamente para capturar su atención y monetizar su comportamiento.

Uno de los fenómenos más preocupantes es la tendencia de los algoritmos a premiar cierto tipo de contenido basado en variables que refuerzan estereotipos de género, raza o belleza. En el caso de Instagram, diversos estudios han demostrado que las imágenes que contienen cuerpos femeninos sexualizados obtienen mayor visibilidad y difusión algorítmica, incluso si el contenido no tiene mayor valor informativo (Cotter, 2021). Este sesgo no es casual: es el resultado de modelos entrenados sobre patrones históricos de comportamiento, donde lo más "engagementable" tiende a ser lo más provocativo.

Además, esto tiene consecuencias psicológicas documentadas. Un estudio de la *Royal Society for Public Health* en Reino Unido clasificó a Instagram como la red más dañina para la salud mental de los jóvenes, asociándola con altos niveles de ansiedad, insatisfacción corporal y depresión, particularmente en mujeres adolescentes (RSPH, 2017). Estos efectos no son incidentales, sino consecuencias estructurales de algoritmos optimizados para resultados comerciales, no para el bienestar humano.

Una característica central de los algoritmos de redes sociales es que son altamente personalizados, pero casi nunca explicables. El “feed” que ve un usuario es resultado de miles de variables analizadas en tiempo real: historial de navegación, tiempo de interacción, likes, comentarios, ubicación, conexiones, etc. Sin embargo, rara vez se informa al usuario por qué está viendo un contenido específico. Esto dificulta cualquier intento de crítica o control por parte del individuo.

TikTok ha sido particularmente señalado por su opacidad. Aunque ha publicado lineamientos sobre cómo funciona su algoritmo, investigaciones periodísticas han mostrado que el contenido puede ser priorizado manualmente por empleados mediante una práctica conocida como “heating,” lo que pone en duda la supuesta neutralidad del sistema (Bloomberg, 2023). Este tipo de intervención encubierta también representa un riesgo ético si se utiliza con fines políticos, ideológicos o comerciales sin conocimiento del usuario.

Diversas encuestas reflejan una creciente desconfianza hacia los sistemas automatizados. Según el *Edelman Trust Barometer 2024*, el 61% de los encuestados a nivel global manifestó no confiar en que las empresas tecnológicas

usen la IA de forma ética. Además, un 67% considera que los gobiernos están rezagados en su capacidad para regular el uso de la inteligencia artificial (Edelman, 2024).

Estos datos coinciden con la creciente demanda de explicabilidad y control. El informe *Global Index on Responsible AI 2024* resalta que menos del 30% de los países analizados tienen marcos legales efectivos para obligar a las empresas a ser transparentes en sus decisiones automatizadas (Global Index, 2024). Esto implica una brecha significativa entre el desarrollo tecnológico y la gobernanza ética.

El uso de algoritmos en redes sociales ilustra con claridad por qué la transparencia y la rendición de cuentas no son solo principios técnicos, sino éticos. Cuando el contenido que moldea nuestras emociones, gustos y decisiones está mediado por sistemas que no podemos comprender ni cuestionar, se comprometen la autonomía individual y el bienestar colectivo.

La responsabilidad no debería recaer únicamente en los usuarios para "entender" los algoritmos, sino en las plataformas para diseñar sistemas comprensibles, auditables y justos. La opacidad no es un defecto colateral, sino una decisión estructural que prioriza la ganancia sobre la equidad. Y en ese contexto, la urgencia de un cambio ético es más evidente que nunca.

La inteligencia artificial ha avanzado rápidamente desde el ámbito de la automatización de tareas simples hasta influir en decisiones complejas con consecuencias significativas para la vida humana. Uno de los dilemas éticos más apremiantes surge al delegar decisiones críticas —como diagnósticos médicos, sentencias judiciales o control vehicular— a sistemas que, aunque potentes en análisis de datos, carecen de empatía, juicio moral y sentido contextual.

En el campo médico, la IA ha mostrado capacidades prometedoras, como la detección temprana de cáncer de piel, retinopatía diabética o anomalías cardíacas. Por ejemplo, un algoritmo desarrollado por Google Health fue capaz de detectar cáncer de mama con mayor precisión que radiólogos humanos en un estudio de 2020 (McKinney et al., 2020). Sin embargo, su implementación masiva sin supervisión clínica ha generado controversias.

El problema radica en que los modelos están entrenados con bases de datos específicas que no siempre reflejan la diversidad global. Estudios han demostrado que algunos algoritmos médicos rinden peor en pacientes de grupos étnicos minoritarios debido a una representación insuficiente en los datos de entrenamiento (Obermeyer et al., 2019). Esto implica que decisiones

médicas podrían ser inexactas o incluso peligrosas para ciertos grupos poblacionales, lo que representa una amenaza directa a la equidad en salud.

Los coches autónomos, como los de Tesla, han sido presentados como el futuro del transporte seguro y eficiente. No obstante, su historia está marcada por accidentes mortales, errores de sistema y dilemas sin resolver. En 2018, una mujer murió atropellada por un vehículo autónomo de Uber en Arizona. La investigación reveló que el sistema había detectado a la peatona, pero no pudo clasificarla correctamente y no activó el freno de emergencia a tiempo (NTSB, 2019).

Casos como este muestran la dificultad de atribuir responsabilidades cuando un error ocurre. ¿Debe responder el fabricante, el desarrollador del software, el dueño del vehículo o el sistema en sí? El vacío legal en torno a la IA autónoma se convierte entonces en un vacío ético: si no hay responsables claros, tampoco hay justicia posible para las víctimas.

La automatización también ha alcanzado el ámbito jurídico. Además del caso de COMPAS ya mencionado, se han desarrollado herramientas como HART (Harm Assessment Risk Tool) en el Reino Unido y PSA (Public Safety Assessment) en varios estados de EE.UU., que buscan predecir la peligrosidad de un detenido. Aunque prometen decisiones más rápidas y "objetivas", los resultados han sido mixtos, y la crítica principal es la misma: falta de explicabilidad y riesgo de perpetuar desigualdades existentes.

Un informe del *AI Now Institute* (2018) advierte que la creciente implementación de IA en el sistema judicial no solo reproduce sesgos estructurales, sino que puede naturalizarlos, al hacerlos parecer "neutrales" por venir de una máquina. La opacidad algorítmica socava principios como la presunción de inocencia, la proporcionalidad y la posibilidad de apelación informada.

Frente a estos desafíos, los marcos regulatorios se encuentran rezagados. La Unión Europea ha sido pionera con su propuesta de **Ley de Inteligencia Artificial** (AI Act), presentada en 2021 y en proceso de aprobación. Esta legislación busca clasificar los sistemas de IA por nivel de riesgo y establecer obligaciones según su impacto potencial. Entre ellas: la prohibición de ciertas aplicaciones invasivas, exigencia de transparencia en sistemas de alto riesgo y derechos para que los ciudadanos reciban explicaciones comprensibles sobre decisiones algorítmicas (European Commission, 2021).

En contraste, países como Estados Unidos y México aún carecen de una regulación integral. En México, por ejemplo, solo se han emitido lineamientos técnicos desde la Agencia Digital de Innovación Pública

(ADIP), sin carácter vinculante. El *Global Index on Responsible AI 2024* ubica a México con un puntaje de 34/100 en capacidades de gobernanza para la IA, lo que indica un nivel incipiente de preparación institucional frente a los desafíos éticos (Global Index, 2024).

Estos casos muestran que la urgencia no solo es conceptual, sino práctica. Si se implementan tecnologías sin un marco claro de supervisión y responsabilidad, el resultado es una normalización de errores que pueden tener consecuencias irreparables. No se trata de detener el progreso, sino de acompañarlo con criterios éticos, jurídicos y sociales que garanticen que la IA beneficie a todos, no solo a quienes la desarrollan.

Volviendo a la pregunta inicial —*¿cuál es la urgencia de exigir transparencia y responsabilidad ética en el desarrollo de algoritmos de inteligencia artificial?*—, la evidencia analizada en este artículo permite afirmar que esa urgencia es inmediata, real y transversal. Desde la justicia penal hasta los diagnósticos médicos, desde el contenido que consumimos en redes hasta los autos que conducimos, los sistemas de IA ya están tomando decisiones que antes eran exclusivamente humanas. Y lo hacen, en muchos casos, sin transparencia, sin supervisión efectiva y sin posibilidad de rendición de cuentas.

La urgencia no es solo por el avance vertiginoso de la tecnología, sino por la ausencia de estructuras legales, educativas y sociales que garanticen que dicho avance sea justo y seguro. La IA no es neutral: reproduce las intenciones, los prejuicios y las prioridades de quienes la diseñan. Y si no se construyen marcos que exijan ética desde su origen, lo que se normalizará será la automatización de la injusticia.

Para enfrentar esta situación, propongo una serie de principios éticos que deberían guiar el diseño, la implementación y la regulación de los sistemas de IA:

1. **Transparencia obligatoria y comprensible:** Toda organización que utilice sistemas de IA con impacto en derechos fundamentales debe estar obligada a explicar, en lenguaje claro y accesible, cómo funcionan sus algoritmos, qué variables utilizan y qué margen de error presentan. Esto no implica revelar secretos industriales, sino garantizar la trazabilidad de decisiones críticas.

2. **Auditorías externas y constantes:** Los sistemas de IA deben ser revisados periódicamente por organismos independientes que evalúen no solo su precisión, sino también su

equidad, impacto social y sesgos latentes. Estas auditorías deben ser públicas y vinculantes.

3. **Participación ciudadana informada:** Es necesario fomentar una cultura digital crítica. Desde la educación básica, se deben enseñar principios básicos sobre cómo funcionan los algoritmos, qué implicaciones tienen y cómo pueden afectar nuestras decisiones. Esta alfabetización algorítmica es clave para una ciudadanía activa y vigilante.

4. **Marcos regulatorios flexibles pero firmes:** Las leyes deben evolucionar al ritmo de la tecnología, pero con principios sólidos que no se negocien: respeto a la dignidad humana, no discriminación, derecho a la explicación, proporcionalidad y reparación en caso de daño.

5. **Diseño ético desde el origen:** La ética no debe ser un control final, sino una parte del proceso de desarrollo tecnológico. Las empresas e instituciones deben integrar equipos multidisciplinarios que incluyan filósofos, sociólogos, juristas y comunidades afectadas en la creación de sistemas de IA.

La urgencia también es cultural. Estamos acostumbrándonos a ver los algoritmos como entidades abstractas, inevitables e incuestionables. Se les atribuye un aura de objetividad técnica que invisibiliza sus fallas y efectos adversos. Pero los algoritmos no son entes mágicos: son construcciones humanas, y como tales, deben estar sujetas al mismo escrutinio, crítica y responsabilidad que cualquier otra herramienta de poder.

El concepto de rendición de cuentas algorítmica implica que toda decisión tomada por una máquina debe poder ser rastreada, explicada y corregida. No basta con decir “el sistema lo decidió”; hay que poder decir *quién lo programó, con qué datos, con qué objetivos y bajo qué criterios éticos*.

Finalmente, esta no es solo una tarea de legisladores o desarrolladores. La ética de la inteligencia artificial es un campo que nos involucra a todos: usuarios, educadores, comunicadores, activistas y estudiantes. Cada clic, cada dato compartido, cada decisión aceptada sin cuestionamiento, contribuye a moldear el tipo de sociedad algorítmica que estamos construyendo.

No se trata de rechazar la tecnología, sino de asumirla con madurez crítica. De exigir que la IA no sea solo eficiente, sino justa. De recordar que el futuro no lo decide un algoritmo: lo decidimos nosotros.

Referencias

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Cotter, K. (2021). "Shadowbanning Is Not a Thing": Black Box Gaslighting and the Power to Define Reality on Social Media. *Social Media + Society*, 7(2). https://doi.org/10.1177/20563051211020268
- Edelman. (2024). *Trust Barometer 2024*. https://www.edelman.com/trust-barometer
- European Commission. (2021). *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. https://ec.europa.eu
- Global Index on Responsible AI. (2024). https://www.global-index.ai
- Hao, K. (2019). Amazon scrapped a secret AI recruiting tool that showed bias against women. *MIT Technology Review*. https://www.technologyreview.com/2018/10/10/139110/amazon-hiring-ai/
- McKinney, S. M., Sieniek, M., Godbole, V., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577, 89–94.
- NTSB. (2019). *Preliminary Report: Highway HWY18MH010*. National Transportation Safety Board. https://www.ntsb.gov
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.
- Royal Society for Public Health. (2017). *#StatusOfMind: Social media and young people's mental health and wellbeing*. https://www.rsph.org.uk
- AI Now Institute. (2018). *AI Now Report 2018*. https://ainowinstitute.org/AI_Now_2018_Report.pdf
- Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1).
- O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing.
- **OpenAI. (2025). ChatGPT (versión GPT-4o) [Modelo de lenguaje]. OpenAI. https://chat.openai.com** *(Utilizado como apoyo para la estructuración argumentativa, redacción , recopilación de fuentes verificables, y referencias en formato APA para el artículo)*

Ética e Inteligencia Artificial

Autor: Manuel Alejandro Hernández Melo

La IA o Inteligencia Artificial se está convirtiendo en una de las tecnologías más revolucionarias que ha creado el ser humano, no solo por la versatilidad de esta o su inimaginable potencial, sino por su explosiva adopción la telaraña de artilugios y estructuras que conforman la vida cotidiana moderna. De acuerdo con datos de Semrush, una reconocida plataforma de análisis de marketing digital utilizada ampliamente por profesionales del sector para obtener métricas precisas de tráfico web y comportamiento de usuarios, el sitio chatgpt.com recibió aproximadamente 5.2 billones de visitas durante el mes de marzo de 2025. Lo que representa una amplia adopción de las IAs dentro del internet y sin considerar otros modelos de IA actualmente disponibles en la red[41]. Incluso dentro de Google, el buscador más utilizado en el planeta, se encuentra incrustada una AI generativa experimental la cual promete suplementar el proceso de búsqueda de información.

Sin embargo y a pesar de la popularidad que se ha generado alrededor de la IA, la velocidad en regulación por parte de gobiernos y el entendimiento detrás de su funcionamiento y consecuencias de uso, por parte de la población en general, no se compara con la aceleración en su desarrollo o implementación en todos los aspectos de la vida. Adicionalmente, el uso de recursos naturales que se invierten en la creación y continuidad de estos avances se presenta cada vez más grande, contrastando el insistente empuje por las grandes compañías tecnológicas por avanzar, inyectar y finalmente beneficiarse económicamente de las IA.

Por ello pregunto, **¿Se debería regular el uso de la inteligencia artificial por parte de la población en general?** Aunque esta nueva cúspide de la imaginación humana tenga una gran capacidad de facilitar nuestras vidas, su uso debería ser reglamentada y controlada directamente por la población, debido a la creciente área gris legal que los estados aún no han logrado reparar.

En este ensayo, se evaluará esta nueva forma de inteligencia analizando argumentos a favor de su regulación: Su

[41] Semrush. (2025, 12 de mayo). *chatgpt.com Website Traffic, Ranking, Analytics [April 2025]*. https://www.semrush.com/website/chatgpt.com/overview/#traffic-journey

costo actual sobre el medio ambiente, puntos actuales de contención en el aspecto de la privacidad, algunas consecuencias de su área gris y las implicaciones éticas que se presentan por la visión a futuro que tiene la empresas que la dirigen; Y sus argumentos en contra: El potencial que presenta para la humanidad, qué impacto tendría una regulación de la IA en las libertades personales, las dificultades de alcanzar su rápido desarrollo y precedentes históricos a la hora de regular tecnologías de este tipo. Remarcando la importancia de construir un futuro con la IA, donde somos conscientes de su costo y consecuencias.

El uso de la inteligencia artificial ha estado en un constante incremento alrededor del mundo, tanto a nivel empresarial como gubernamental. De acuerdo con el "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value" de la consultora estratégica global McKinsey, el 72% de las organizaciones consultadas ocupan IA generativa de texto en alguna función[42]; Por el otro lado en el ámbito Gubernamental en 2025 se presentaron usos de IA para el manejo de datos de la IRS y generación de arte con propósito propagandístico en los Estados Unidos. De igual manera, la adopción de esta tecnología por la población en general continúa creciendo, según el informe de una encuesta mundial llevado a cabo por la Universidad de Toronto en 2024 y citado por la Universidad de Stanford señala que "un 63% de los encuestados están familiarizados con ChatGPT, y casi la mitad de ellos lo usan al menos una vez por semana". Reflejando su creciente integración el positivismo por las IAs va en crecimiento con las poblaciones más jóvenes y de mayor nivel educativo y económico opinando positivamente en múltiples estudios.

En términos de desarrollo y concentración de uso de la IA, el *AI Index Report 2024* identifica a los Estados Unidos como el centro de la innovación, inversión privada y uso de inteligencias artificiales, seguido por la República Popular China, que se encuentra acortando la distancia entre sí y los estados Unidos[43]. Otros gigantes que han aumentado su interés en esta tecnología, son India y el Reino Unido también se presentan como países con emergentes intereses económicos, de investigación y

[42] Singla, A., Sukharevsky, A., Yee, L., & Chui, M. (2024, May 30). *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value*. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-202 4

[43] Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. (2024). *The AI Index 2024 Annual Report*. AI Index Steering Committee, Institute for Human-Centered Artificial Intelligence, Stanford University. https://hai.stanford.edu/ai-index/2024-ai-index-report

de consumo relacionado con la IA. Respecto al uso de esta herramienta a un nivel personal, según el informe de Google y Ipsos más del 40% de los usuarios que usan el internet reportan haber usado la IA [4]. Entre los usos más comunes de la IA entre la población en general se encuentran: La generación de contenido con propósitos de entretenimiento (imágenes o texto), apoyo en actividades educativas y apoyo en tareas del día a día cómo redacción, traducción y creación de material visual. Como señala en informe de Google y Ipsos "La IA se está convirtiendo en una herramienta contiendan que permite a las persona ahorrar tiempo, aprender nuevas habilidades y expresarse creativamente"[44].

Sin embargo, con la creciente adopción de la IA alrededor del mundo su desarrollo se sigue viendo concentrado en conglomerados económicos concentrados alrededor de las principales potencias mundiales. Adicionalmente, aunque su uso está en crecimiento alrededor del globo los nexos donde se concentra su uso se ven limitados a las áreas más desarrolladas, con acceso a internet y posiciones privilegiadas en la sociedad expandiendo la creciente brecha de desigualdad no sólo entre naciones, sino también en comunidades. Por otro lado, enfocándonos en el uso que se le da a la herramienta, se tiene que ponderar frente al enorme área gris que existe entre las capacidades de la IA y la ley, es un crimen generar una imagen inapropiada de alguien con IA o generar material falso con fines maliciosos. De la misma manera, se puede justificar los costos al medio ambiente en forma de agua, minerales y polución del aire, que afecta a las poblaciones más vulnerables por generar un simple meme o una ayuda en una tarea escolar.

En relación con lo anterior, los gobiernos de los países alrededor del mundo han comenzado a generar legislaciones, estatutos y mociones centradas en la IA, más específicamente alrededor de su desarrollo y uso. Sin embargo, debido al tiempo que toma diseñar leyes, revisarlas y tanto identificar como resolver huecos dentro de estas, en contraste con la velocidad con la que avanza la tecnología, el poder jurídico no ha logrado delimitar correctamente a la IA dentro del marco de la Ley. Consecuentemente y sumado a su popularidad, facilidad de uso y poca transparencia de datos por parte de las empresas, es natural que la IA se haya adoptado como una herramienta para facilitar e incluso generar nuevas formas

[44] Azevedo Lohr, A. (2025, 14 de enero). *Google / Ipsos Multi-Country AI Survey 2025*. Ipsos. https://www.ipsos.com/en-us/google-ipsos-multi-country-ai-survey-2025

de criminalidad[45].

En términos éticos o incluso legales la mayoría de las Inteligencias Artificiales generativas como chat GPT se encuentran reguladas o moderadas con apenas pocos filtros y controles agregados que impiden que la herramienta sea usada inapropiadamente según su diseñador. Aunque estos filtros pueden ser sobrepasados (por ejemplo, mediante la técnica de "en un caso hipotético") y en los peores casos no cumplen un propósito más allá de limitar el acceso a información sensible. Un caso aplicable es DeepSeek, un motor con IA que censura eventos como la masacre de la Plaza de Tiananmen.

Los ejemplos más relevantes sobre el uso inapropiado de la Inteligencia Artificial y aunque no ético, teóricamente legal dentro del área gris que se agranda aún más entre la tecnología y la ley, son los siguientes: Suplantación de figuras políticas como el caso de la princesa Leonor para estafas digitales[46], su uso por organizaciones criminales para generar imágenes y audios “deepfake” para la extorsión[47] y para difamación de figuras públicas como la denuncia ante la generación de contenido inapropiado de su imagen por la senadora Andrea Chávez[48]. El caso más infame se presenta a finales del año 2024, donde el alumno Diego “N” del Instituto Politécnico Nacional fue puesto en juicio por la posesión y generación de material pornográfico generado con IA a partir de imágenes utilizadas sin consentimiento de sus compañeras, pero fue absuelto al argumentar la falta de evidencia contundente según el juez[49]. No solo demostrando incapacidad del gobierno mexicano por adaptarse ante la IA, pero marcando un precedente alarmante o como lo puso la senadora Larios Pérez ante el Congreso de la Ciudad de México:

“Con esta decisión el juez sentó un precedente histórico de negligencia, de

[45] Global AI Ethics and Governance Observatory. https://www.unesco.org/ethics-ai/es/mexico

[46] Cantó, P. (2024, diciembre 3). *“Creí que hablaba con Leonor y ahora estoy endeudada”: así suplantan a la Princesa de Asturias para realizar estafas en Latinoamérica*. *El País*.https://elpais.com/tecnologia/2024-12-03/crei-que-hablaba-con-leonor-y-ahora-estoy-endeudada-asi-suplantan-a-la-princesa-de-asturias-para-realizar-estafas-en-latino america.htmlEl País+2El País+2El País+2

[47] InSight Crime. (2024, enero 12). *Cuatro formas en que la inteligencia artificial está transformando el crimen organizado en América Latina*. https://insightcrime.org/es/noticias/cuatro-formas-inteligencia-artificial-transformando-crimen-organizado-america-latina/

[48] Jiménez Urzúa, L. (2024, febrero 7). *El caso de Andrea Chávez: violencia digital y la necesidad de justicia para todas*. *El Universal*.

[49] Congreso de la Ciudad de México. (2024, diciembre 6). *Congreso solicita a juez validar pruebas en caso de violencia sexual digital*. https://www.congresocdmx.gob.mx/comsoc-congreso-solicita-juez-validar-pruebas-caso-violencia-sexual-digital-5855-1.html

opacidad, de complicidad y lo más grave de impunidad en contra de las mujeres, dejando abierta la puerta a que más agresores utilicen la tecnología como herramienta de opresión, sabiendo que el sistema los puede proteger, y al mismo tiempo cerrándole la puerta a la justicia a las ocho estudiantes, mujeres jóvenes valientes, que están luchando, alzando la voz para que este crimen no quede impune" (Senadora Larios Pérez, Congreso de la Ciudad de México, 2024)

Este caso no sólo demuestra la incapacidad del gobierno mexicano por adaptarse ante la IA, sino que también sirve como una advertencia sobre los riesgos de permitir amplias áreas grises entre la supervisión gubernamental y los avances tecnológicos.

Como respuesta a esto, la UNESCO género una carta de recomendaciones específicamente sobre la gobernanza de las inteligencias artificiales, enfatizando la urgencia por establecer principios vinculantes que protejan los derechos humanos. El documento hace un punto especial sobre la falta de rendición de cuentas, la exclusión de grupos minoritarios y de bajos recursos en el diseño y regulación de la tecnología y la necesidad de establecer marcos normativos centrados en la equidad, transparencia y supervisión por parte de la población en general[50]. Sin embargo, aunque este llamado internacional refuerza la urgencia de actuar ante estos avances, desde su adopción en 2021 no ha tenido un gran impacto sobre la emergente crisis.

Asimismo, el Global Index on Responsible IA ha demostrado las titánicas disparidades en la adopción de IA comparado con la implementación de políticas públicas sobre inteligencia artificial responsable, destacando la poca previsión por parte de la mayoría de los países del mundo[51]. El índice señala que, los mecanismos de vigilancia actuales y las estructuras legales están considerablemente rezagadas frente al acelerado desarrollo y adopción de esta tecnología, hecho aparente por las aisladas iniciativas relacionadas con IA. Lo que resulta en una incompatibilidad entre la innovación y regulación que provee un espacio de crecimiento a los riesgos que presenta la IA en escenarios como son la educación, justicia y seguridad, exponiendo a la población ante un arma libre de supervisión y con potencial en constante crecimiento para la violencia digital.

[50] UNESCO. (2021). *Recomendación sobre la ética de la inteligencia artificial*. https://unesdoc.unesco.org/ark:/48223/pf0000385082

[51] Global Center on AI Governance. (2024). *The Global Index on Responsible AI*. https://www.global-index.ai/

Paralelo a los problemas emergentes legales, también surge la preocupación por el costo ambiental que se relaciona con la inteligencia artificial, en razón de que requiere de una enorme inversión inicial para su desarrollo. Siendo dividida en capital monetario, materias primas como el litio y un consumo intensivo de energía y agua para su entrenamiento[52]. Según investigadores de la universidad de Massachusetts, en 2019 la huella de carbono considerando para entrenar un solo modelo grande de PLN (Procesamiento de lenguaje natural) es 626,155 libras de CO2 o 284,02 toneladas de CO2[53]. Representando una alarma a ser considerada, al evaluar el interés de la población en general por adoptar en diferentes ámbitos de su vida a modelos incluso más modernos e intensivos en su generación de huella de carbono.

Sin embargo, se puede declarar que los individuos no tienen un poder directo en decidir cómo se desarrolla, cuál es su objetivo o que se invierte en el desarrollo de este tipo de modelos. Pero es indudable que el uso de los modelos de IA más populares como Chat GPT en la mayoría de los casos es totalmente voluntaria, aunque influenciada por presiones sociales, las empresas interesadas en el consumo y la indiferencia del gobierno ante las implicaciones de su uso. Relacionado con esto la huella de carbono/costos energéticos de escribir una oración dentro del modelo Chat GPT-3 conlleva un consumo energético estimado de aproximadamente 0.003 kWh y entre 0.1 y 0.5 litros de agua dulce para el enfriamiento, en el centro de datos promedio. Lo que exacerba los impactos tanto directos, en forma de sequías y apagones en comunidades adyacentes causados por el consumo de centros de información y en los desastres relacionados con el calentamiento global en general. Un ejemplo puede ser visto en el estado de Querétaro, México, la rápida expansión de centros de información de empresas como Amazon, Google y Microsoft han causado presiones considerables en los recursos hídricos red eléctrica en una región afectada por sequías frecuentes y infraestructura limitada[54].

El desarrollo y la implementación de la Inteligencia Artificial están dominados

[52] Aguilar, A. & El Sol de México. (2023, May 24). *¿La Inteligencia Artificial afecta al medio ambiente?* NewsBankinc. Retrieved May 15, 2025, from https://infoweb-newsbank-com.us1.proxy.openathens.net/apps/news/document-view?p=AWNB&docref=news/191B81D829462D68

[53] Araiz Huarte, D. E. (2023, 16 de enero). La inteligencia artificial como agente contaminante: concepto jurídico, impacto ambiental y futura regulación. *Actualidad Jurídica Ambiental*, (130). https://doi.org/10.56398/ajacieda.00071

[54] Graham, T. (2024, septiembre 25). *Mexico's datacentre industry is booming – but are more drought and blackouts the price communities must pay?* The Guardian. https://www.theguardian.com/global-development/2024/sep/25/mexico-datacentre-amazon-google-queretaro-water-electricity

por un número reducido de corporaciones tecnológicas que concentran los recursos clave como datos, infraestructura computacional, talento especializado y capital. Esta concentración de poder ha posicionado al sector privado como el maestro de la IA, desplazado al sector académico y organismos públicos, lo que ha limitado el enfoque de la tecnología a objetivos de lucro, encargando la supervisión de esta tecnología bajo el concepto "Autorregulación" dentro de las mismas empresas. Adicionalmente, la falta de regulación efectiva y velocidad de avance tecnológico ha generado vacíos legales entre los marcos legales existentes, que se ha aprovechado por las empresas. Como resultado, la población en general no ha sido consultada adecuadamente sobre la dirección del desarrollo tecnológico no sobre las áreas de su vida cotidiana donde se incrusta sin previo aviso. Los individuos son tratados como cifras de donde se recolecta información, hacia donde se vende un servicio y a la cual en términos de aprobación se le presentan sólo dos opciones: adoptar la Inteligencia Artificial o ser dejado atrás por la competencia. Una dinámica que refleja una falta profunda de consentimiento informado y de transparencia a la hora de implementar tecnologías de gran potencial para transformar la sociedad.

Adicionalmente, las empresas tecnológicas han promovido la adopción de la IA mediante estrategias de mercadotecnia que destacan su potencial de facilitar la vida, en términos de eficiencia y personalización, mientras que minimizan, ofuscan e incluso ignoran los costos asociados. Estas estrategias incluyen la presentación de servicios aparentemente gratuitos que, en realidad, implican la recopilación intensiva de datos personales y la falta de transparencia en las políticas de privacidad. Otro ejemplo es la adición de inteligencia artificial en el motor de búsqueda más usado del mundo Google, donde con el paso del tiempo han potenciado las herramientas de búsqueda y clasificación con IA e incluso se ha adherido una IA generativa de texto experimental llamada Visión. Finalmente, y muy preocupantemente considerando la huella de carbono de la IA, las empresas que desarrollan esta tecnología han optado por omitir y minimizar con estimaciones en reportes y presentaciones públicas los costos ambientales relacionados con el desarrollo y uso de la IA, con muchos reportes académicos reportando esta información siendo apenas estimaciones tomadas desde el exterior.

A pesar de que, las preocupaciones directamente anexas al mal uso, implicaciones éticas y de seguridad en relación con la Inteligencia Artificial son válidas, también es fundamental

considerar los riesgos asociados con una regulación que limite el acceso de estas nuevas tecnologías a la población en general. Asimismo y reconociendo todo el panorama, se puede asumir con seguridad que las nuevas normas tendrían el potencial de obstaculizar el desarrollo tecnológico de la IA. Lo que tendría repercusiones en la población al limitar los beneficios de esta tecnología, áreas de innovación donde puedan participar, la aparente equidad en el acceso de uso y el potencial transformador que presenta. Por ello en esta sección se explorarán los argumentos en contra de la regulación del uso de la IA por la población en general.

Uno de los argumentos más destacados que se presentan en contra de una regulación estrictamente sobre el uso de la Inteligencia Artificial, por parte de la población en general, se basa en su potencial de transformar la vida humana hacia un mejor futuro. La IA se presenta como una herramienta que amplifica las capacidades y actividades humanas, al optimizar la creatividad, productividad y el acceso a la amplia librería de información que es el internet, demostrando como un agregado invaluable desde la educación hasta el emprendimiento digital[55]. Autores como Brynjolfsson y McAfee (2014) han señalado en su libro "The Second Machine Age", que la IA tiene el poder de empoderar a los individuos al democratizar el acceso a capacidades antes reservadas a expertos o grandes corporaciones. Como afirman: *"Technology is not destiny. We shape our destiny. We shape our technologies, and our technologies shape us"* (Brynjolfsson & McAfee, 2014, p. 10). Adyacente a ello, restringir el uso de este gran avance podría limitar las oportunidades para que poblaciones con menos recursos (pero aún con acceso a internet y electricidad) o educativos aprovechen para ocupar esta herramienta para mejorar su calidad de vida, reduciendo las brechas entre la sociedad al proveer nuevas formas de expresión.

Además, la imposición de controles generalizados sobre el uso cotidiano de la IA podría inhibir la innovación desde abajo, o en otras palabras, la que surge a partir del uso informal y creativo por los usuarios comunes ajenos a los desarrolladores. Como menciona Erick Von Hippel en Democratizing Innovation, sostiene que "users are the first to develop many and perhaps most new industrial and consumer products" (Von Hippel, 2005,p. 1)[56], y subraya que "innovation by users has become a major force" (Von Hippel, 2005, p. 2. Limitar el

[55] Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies* [E-book version].

[56] Hippel, E. von. (2005). *Democratizing innovation / Eric von Hippel*. MIT Press. https://directory.doabooks.org/handle/20.500.12854/77862

acceso generalizado de las inteligencias artificiales, tendría el impacto de sofocar esta dinámica de descubrimiento colectivo, restringiendo el desarrollo de avances en tecnología o bien soluciones menores que tiene gran impacto a niveles de comunidad. Consecuentemente, una regulación excesiva no sólo presenta riesgo de excluir a sectores vulnerables de la población, cortándoles el aprovechamiento de la herramienta, sino también atasca los procesos de innovación inclusiva, descentralizada y democrática.

Un segundo punto crítico por evaluar en el debate sobre la regulación del uso de la inteligencia artificial por parte de la población general se relaciona con su función como herramienta de libre expresión. Actualmente, los principales organismos directamente que filtran el uso de la IA son los mismos desarrolladores y las empresas detrás de ellos. Lo que ha resultado, de manera simplificada, en que el usuario tenga un considerable grado de libertad a la hora de decidir cómo ocupar la herramienta. Llevando a una amplia versatilidad de uso y proveyendo al usuario con una avenida para expandir sus formas de expresión, al proveerles la habilidad de generar material escrito, auditivo e incluso visual con solo un “prompt”. En este sentido, la IA se ha convertido para muchos en una extensión de sus derechos de libertad de expresión, al posibilitar nuevas formas de comunicación y creación, aunque como se menciona en las anteriores secciones ha llevado a considerables repercusiones.

No obstante, trasladar la capacidad regulatoria de las empresas al gobierno plantea desafíos significativos y 2 específicamente relacionados con la libertad de expresión. Debido a que conlleva una amplia participación y consentimiento por parte de la población[57]. La posibilidad de que no se considere la participación pública de todos los sectores de la sociedad o simplemente se asuman a través de los cuerpos reguladores, podría desembocar en formas de censura o supresión del discurso. Ejemplos como los modelos de generación de texto censuran información sensible y la utilización de IA en sistemas de vigilancia en China, particularmente en el caso de la minoría Uigur, revelan cómo la tecnología puede ser empleada como un mecanismo de represión en contextos autoritarios. Paralelo a esta dificultad, el entendimiento multidisciplinario sobre el funcionamiento, tipos de uso y potencial que se relacionan con la IA son esenciales. Ya que al ser una herramienta multifacética podría caer en

[57] Crawford, K. (2021). *The Atlas of AI : Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. https://doi.org/10.2307/j.ctv1ghv45t

ambigüedades que permitan su abuso o limiten usos legítimos o creativos. Como advierte Nemitz en su libro Constitutional democracy and technology in the age of artificial intelligence (2018), regular la IA requiere de una comprensión transversal, que combine conocimientos legales, técnicos, éticos y sociales para evitar legislaciones excesivamente restrictivas o mal enfocadas.

Sumado a los riesgos relacionados con la libertad de expresión, la velocidad del desarrollo tecnológico en inteligencias artificiales es una realidad que considerar, pues cualquier intento de regulación a largo plazo necesita tener una estructura suficientemente flexible para abarcar todo lo necesario y suficientemente sólida para ser aplicable. El ritmo con el que emergen nuevas aplicaciones, modelos y funcionalidades tiende a dejar obsoletos a los marcos regulatorios incluso cuando estos apenas están por ser implementados, como visto con legislaciones en la Unión Europea. La "brecha de velocidad" entre el avance tecnológico y la formulación de políticas públicas no es un fenómeno nuevo, pero como señala ampliamente la literatura la IA presenta un reto especialmente grande alcanzar un balance apropiado. Como se explica en "El uso Malicioso de la Inteligencia artificial" (2018) , *"the lifecycle of innovation in AI is much shorter than that of regulation, which creates a mismatch between when a risk emerges and when institutions are able to respond to it"* (*The Malicious Use of Artificial Intelligence*, p. 45).

En consecuencia, lograr un punto de equilibrio a largo plazo entre la prevención de usos maliciosos y la protección de libertades individuales resulta casi inalcanzable. Las regulaciones considerablemente inflexibles o específicas tienden a la obsolescencia debido a los cambios constantes, mientras que un enfoque más amplio o general tiene el riesgo de ser ineficaz al presentar huecos legales. Por ello es de suma importancia que la población y especialmente expertos en IA participen en tiempo real, ya que la única forma de regular la IA es con un enfoque adaptativo e incluso predictivo. Convirtiendo a la retroalimentación activa de los ciudadanos, la evaluación de algoritmos usados por IA y transparencia gubernamental como imperativos para preservar los principios fundamentales de la democracia en un mundo que hace su transición hacia una nueva era.

La Inteligencia Artificial se está convirtiendo en uno de los avances tecnológicos con el mayor potencial de transformar la vida contemporánea, proveyendo múltiples herramientas para facilitar el progreso personal y colectivo. Paralelamente, se han levantado múltiples problemáticas alrededor de las implicaciones éticas que conlleva su uso,

sociales a la hora de considerar la equidad y de costo ambiental. En este ensayo se han presentado los riesgos y costos asociados con el uso irrestricto de la IA: el impacto ambiental, potencial que tiene de agrandar desigualdades, las amenazas a la privacidad, el vacío legal y la concentración de poder en manos de corporaciones tecnológicas. Al igual que las dificultades relacionadas con una posible regulación: excusa para la censura, Legislación demasiado específica o generalizada, inhibir participación de la población en general y precedente histórico.

Ante este panorama, se tiene que encontrar un equilibrio entre la tecnofobia y la adoración ciega de una herramienta. Por ello es imperativo construir una vía intermedia para su uso por la población en general: basada en regulación bajo un marco ético, con una participación activa de la población en general y transparente por parte de los gobiernos, para no limitar de manera arbitraria a la sociedad, pero para salvaguardarla. Regular la Inteligencia Artificial no es censurarla, pero es reconocer las consecuencias relacionadas con su uso masivo y su rápido desarrollo requieren responsabilidad colectiva y no solo por parte de quienes lo diseñan y monetizan. En conclusión, el objetivo no es detener el progreso, sino tener presente sus costos y consecuencias, para así guiarlo hacia un futuro donde la IA sirve a la humanidad sin amenazar su bienestar.

Referencias

- Semrush. (2025, 12 de mayo). *chatgpt.com Website Traffic, Ranking, Analytics [April 2025]*. https://www.semrush.com/website/chatgpt.com/overview/#traffic-journey
- Singla, A., Sukharevsky, A., Yee, L., & Chui, M. (2024, May 30). *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value*. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-202 4
- Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. (2024). *The AI Index 2024 Annual Report*. AI Index Steering Committee, Institute for Human-Centered Artificial Intelligence, Stanford University. https://hai.stanford.edu/ai-index/2024-ai-index-report
- Azevedo Lohr, A. (2025, 14 de enero). *Google / Ipsos Multi-Country AI Survey 2025*. Ipsos. https://www.ipsos.com/en-us/google-ipsos-multi-country-ai-survey-2025 [5]
- *México*. (2024, July 30). Global AI Ethics and Governance Observatory.

https://www.unesco.org/ethics-ai/es/mexico

- Cantó, P. (2024, diciembre 3). *"Creí que hablaba con Leonor y ahora estoy endeudada": así suplantan a la Princesa de Asturias para realizar estafas en Latinoamérica*. *El país.* https://elpais.com/tecnologia/2024-12-03/crei-que-hablaba-con-leonor-y-ahora-estoy-endeudada-asi-suplantan-a-la-princesa-de-asturias-para-realizar-estafas-en-latino america.htmlEl País+2El País+2El País+2
- InSight Crime. (2024, enero 12). *Cuatro formas en que la inteligencia artificial está transformando el crimen organizado en América Latina*.

  https://insightcrime.org/es/noticias/cuatro-formas-inteligencia-artificial-transformand o-crimen-organizado-america-latina/
- Jiménez Urzúa, L. (2024, febrero 7). *El caso de Andrea Chávez: violencia digital y la necesidad de justicia para todas*. *El Universal*. https://www.eluniversal.com.mx/opinion/leslie-jimenez-urzua/el-caso-de-andrea-chav ez-violencia-digital-y-la-necesidad-de-justicia-para-todas/
- Congreso de la Ciudad de México. (2024, diciembre 6). *Congreso solicita a juez validar pruebas en caso de violencia sexual digital*.

  https://www.congresocdmx.gob.mx/comsoc-congreso-solicita-juez-validar-pruebas-ca so-violencia-sexual-digital-5855-1.html
- UNESCO. (2021). *Recomendación sobre la ética de la inteligencia artificial*. https://unesdoc.unesco.org/ark:/48223/pf0000385082
- Global Center on AI Governance. (2024). *The Global Index on Responsible AI*. https://www.global-index.ai/
- Aguilar, A. & El Sol de México. (2023, May 24). *¿La Inteligencia Artificial afecta al medio ambiente?* NewsBankinc. Retrieved May 15, 2025, from https://infoweb-newsbank-com.us1.proxy.openathens.net/apps/news/document-view? p=AWNB&docref=news/191B81D829462D68
- Araiz Huarte, D. E. (2023, 16 de enero). La inteligencia artificial como agente contaminante: concepto jurídico, impacto ambiental y futura regulación. *Actualidad Jurídica Ambiental*, (130). https://doi.org/10.56398/ajacieda.00071
- Lehrer, M., & Stark, D. (2025). *Smoke and mirrors? AI regulation and corporate power*. *University of Illinois Journal of Law, Technology & Policy*, 2025(1).

  https://dx.doi.org/10.2139/ssrn.4736131
- Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies* [E-book version].
- Hippel, E. von. (2005). *Democratizing innovation / Eric von Hippel*. MIT Press.

https://directory.doabooks.org/handle/20.500.12854/77862

- Crawford, K. (2021). *The Atlas of AI : Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. https://doi.org/10.2307/j.ctv1ghv45t
- Mozur, P., Kessel, J. M., & Chan, M. (2019). *Made in China, Exported to the World: The Surveillance State*. The New York Times.

  https://www.nytimes.com/2019/04/24/technology/ecuador-surveillance-cameras-polic e-government.html
- Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, *376*(2133), 1–14.
- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., - - - Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*.
- Graham, T. (2024, septiembre 25). *Mexico's datacentre industry is booming – but are more drought and blackouts the price communities must pay?* The Guardian. https://www.theguardian.com/global-development/2024/sep/25/mexico-datacentre-am azon-google-queretaro-water-electricity

Reflejos de poder: políticas de una sociedad fragmentada

Autora: Margarita Córdoba Zepeda

La ética algorítmica frente a la distorsión de la justicia social

Al algoritmo hay que verlo como un espejo: no crea nada desde cero, solo devuelve una imagen de aquello que le enseñamos. Pero si la luz que lo ilumina está torcida y si la historia que alimenta sus cálculos está plagada de prejuicios, lo que devuelve no es claridad, sino distorsión. La inteligencia artificial ha llegado a ocupar un rol central en las decisiones públicas: en hospitales, juzgados, escuelas, oficinas de migración. Sin embargo, como advierte Virginia Eubanks en *Automating Inequality (2018)*, estos sistemas no corrigen nuestras fallas. Las aprenden, las refinan y las aplican con una eficacia tan limpia que puede parecer neutral.

Detrás de cada algoritmo público hay decisiones políticas ocultas tras fórmulas matemáticas. Y cuando esas decisiones se aplican sin apelación, sin explicación, sin rostro, su impacto se vuelve brutal. En Detroit, el sistema automatizado de valoración fiscal sobrestimó el valor de las viviendas en barrios pobres por años, provocando desahucios masivos que afectaron sobre todo a comunidades afroamericanas. En el Reino Unido, un sistema de predicción de abandono escolar etiquetó a estudiantes de bajos ingresos como “casos perdidos”, reduciendo la atención que se les brindaba en lugar de aumentarla. En Brasil, un modelo de evaluación de beneficiarios de ayudas sociales redujo los pagos a miles de familias con base en indicadores que nunca fueron explicados públicamente. No son errores técnicos. Son decisiones políticas automatizadas.

Eubanks lo resume con claridad: la automatización en lo público no es inocente. La “casa pobre digital” que ella describe es el resultado de siglos de desigualdad traducidos en datos, y luego reinterpretados como diagnósticos supuestamente objetivos. La eficiencia se vuelve un valor supremo, y con ello se justifican recortes, exclusiones, penalizaciones. La IA no es un oráculo: es una estructura de poder revestida de algoritmos.

Y entonces, el dilema ya no puede presentarse como una cuestión meramente técnica. ¿Qué clase de justicia es posible cuando las decisiones más sensibles de quién merece apoyo, quién representa un riesgo, quién es

confiable, se basan en patrones extraídos de un pasado injusto?
¿Qué tipo de futuro construimos si aceptamos que la máquina, por ser rápida y precisa, merece más confianza que el juicio humano?

Aquí es donde la reflexión debe hacerse más profunda, más incómoda y más urgente. No estamos ante un problema de algoritmos defectuosos o de sesgos técnicos que una actualización pueda corregir. La inteligencia artificial es un terreno de disputa sobre el poder y la justicia, un espejo que no sólo refleja sino que también moldea las estructuras sociales. La aparente objetividad de sus cálculos no es un refugio seguro, sino una máscara que oculta decisiones políticas y éticas tomadas en la sombra. Con ello, cada dato que alimenta estos sistemas es una huella de exclusión, marginación o estigmatización histórica.

Esta no es una cuestión meramente técnica. Las decisiones automatizadas que influyen en quién merece apoyo, quién representa un riesgo, o quién es confiable, se sustentan en fundamentos éticos que deben ser examinados con rigor. En este escenario, la eficiencia algorítmica emerge como la promesa seductora de maximizar el bienestar colectivo, una suerte de panacea utilitarista para sociedades complejas y saturadas de información. Desde esta óptica, la lógica utilitarista, fundada en las ideas de Jeremy Bentham y John Stuart Mill, justifica la aplicación de la inteligencia artificial en la gestión pública como un medio para producir el mayor bien para el mayor número.

La calculadora fría del algoritmo ofrece una optimización de recursos, reducción de errores humanos y un alcance más amplio en la toma de decisiones públicas, operando con una precisión que escapa a la subjetividad humana. Como señala Mill en *Utilitarianism (1863)*, “la felicidad es el único fin al que deben dirigirse las acciones humanas”, y en ese sentido, la IA se presenta como un vehículo para acelerar ese fin en un mundo donde la complejidad parece inabarcable.

Sin embargo, esta defensa utilitarista de la eficiencia debe ser interrogada en sus propios términos. Maximizar la felicidad no es sinónimo de justicia ni respeto a la dignidad individual. La lógica utilitarista, por más eficiente que sea, puede sacrificar a minorías o individuos en aras del bienestar mayoritario. El sistema que etiquetó a estudiantes pobres en el Reino Unido como “casos perdidos” ilustra cómo esta búsqueda de eficiencia puede derivar en exclusión y deshumanización, reduciendo el acceso a bienes fundamentales como la educación.

Frente a este utilitarismo frío, la ética de la virtud (*Aristóteles*) invita a un replanteamiento profundo. Alasdair MacIntyre nos recuerda que la moral no reside solo en los resultados, sino en el

carácter y las intenciones que guían las acciones. Virtudes como la justicia, la prudencia y la empatía son difíciles, si no imposibles, de codificar en un algoritmo, pero indispensables para construir una comunidad humana que no se reduzca a cifras o estadísticas. La tecnología, aunque poderosa, no es ni puede ser virtuosa si no está imbuida de una intención ética clara y de responsabilidad compartida.

La tensión entre eficiencia y virtud es una realidad tangible: los algoritmos, lejos de ser objetos neutrales, están inmersos en contextos sociales y políticos que determinan su impacto. Ignorar la complejidad y diversidad de las vidas humanas que evalúan es condenar estas herramientas a reproducir prejuicios históricos bajo el disfraz de la objetividad. Si la modernidad ha olvidado la importancia de las prácticas humanas en la formación del carácter y la comunidad, entonces no puede esperar que la tecnología supla esa ausencia.

En las discusiones contemporáneas sobre el papel de la inteligencia artificial en la esfera pública, una defensa frecuente se ancla en la búsqueda de la eficiencia como principio rector. Desde este punto de vista, la incorporación de sistemas algorítmicos se presenta como un salto cualitativo hacia la optimización del bienestar social, donde la justicia se mide en términos de resultados cuantificables y utilidad máxima. Esta justificación se fundamenta primordialmente en el utilitarismo, que sostiene que la acción correcta es aquella que produce la mayor felicidad para el mayor número de personas.

El utilitarismo se convierte en un lenguaje cómodo para quienes diseñan y defienden sistemas automatizados, pues ofrece una justificación matemática a las decisiones públicas. La idea de que los algoritmos pueden analizar grandes volúmenes de datos y proyectar consecuencias optimizadas hace que la eficiencia algorítmica se presente como un ideal indiscutible: minimizar costos, reducir errores humanos, agilizar procesos y extender la cobertura de servicios. John Stuart Mill escribía en *Utilitarianism* (1863) que “el desinterés en los resultados particulares de nuestras acciones, siempre que aumenten la felicidad general, es una expresión de moralidad madura y racional”, y es precisamente esta promesa de racionalidad la que seduce a los tecnócratas y políticos contemporáneos.

Sin embargo, esta exaltación de la eficiencia no es una defensa simplista, sino que, en su formulación más matizada, incluye el reconocimiento de que la automatización debe responder a un objetivo colectivo que trascienda intereses particulares. Bajo esta luz, la IA es vista como una herramienta capaz de materializar el ideal ilustrado del progreso: la mejora tangible de las

condiciones materiales de vida. En la administración pública, esto puede traducirse en sistemas de salud que detectan con anticipación riesgos médicos, programas sociales que identifican beneficiarios con mayor precisión, o procesos judiciales que reducen la discrecionalidad humana y el error.

Tomemos el ejemplo del sistema predictivo de salud utilizado en ciertos hospitales, que a partir de datos biométricos y antecedentes clínicos, anticipa la necesidad de intervenciones preventivas, mejorando los índices de supervivencia. Desde una mirada utilitarista, esta aplicación es ejemplar, pues maximiza el bienestar colectivo mediante el ahorro de recursos y la atención oportuna. Como apunta Peter Singer, un filósofo contemporáneo que se adscribe al utilitarismo, "la moralidad exige que busquemos siempre la manera más eficaz de mejorar la vida de las personas", y en este sentido, la IA es una herramienta ética cuando su uso aumenta la suma total de felicidad y bienestar.

No obstante, dentro de esta defensa también aparece la teoría de la virtud, que aunque menos discutida en el debate tecnológico, aporta un enfoque crucial para comprender las limitaciones éticas de la eficiencia algorítmica. La ética de la virtud, retomada en la filosofía contemporánea por Alasdair MacIntyre (1984) y Philippa Foot (2001), centra la atención en el carácter moral del agente y las cualidades que constituyen una vida buena. En lugar de focalizarse exclusivamente en resultados o reglas, la virtud enfatiza la prudencia, la justicia, la templanza y la empatía como condiciones necesarias para actuar correctamente.

Aplicada a la inteligencia artificial, esta perspectiva advierte que la eficiencia sin virtud puede ser una vía directa a la deshumanización. Un algoritmo puede optimizar procesos, pero no puede incorporar compasión ni discernimiento moral auténtico. No puede, por ejemplo, comprender las circunstancias particulares de un individuo que sufre una exclusión injusta, ni valorar la complejidad de una vida humana reducida a un perfil de datos. Como observa MacIntyre en *After Virtue* (1984), "la modernidad ha perdido el sentido de comunidad moral", y los sistemas algorítmicos corren el riesgo de perpetuar esta fractura al tratar a las personas como casos o cifras.

Un ejemplo claro se encuentra en los sistemas de evaluación social en Brasil, donde la reducción mecánica de pagos a familias vulnerables, basada en indicadores no transparentes, no solo produjo injusticias concretas, sino que también erosionó la confianza ciudadana en las instituciones. Desde una óptica virtuosa, este fenómeno no solo es un

fallo técnico, sino una expresión de la ausencia de justicia y prudencia en el diseño y aplicación de estas tecnologías.

Este contraste entre utilitarismo y ética de la virtud resalta una paradoja central: la eficacia técnica de la IA puede coincidir con un déficit moral profundo. La automatización puede funcionar perfectamente dentro de sus parámetros, pero si esos parámetros están diseñados sin una orientación ética que valore la dignidad humana más allá de la suma de utilidades, el resultado será un sistema eficiente pero injusto. La virtud nos recuerda que la justicia no es solo cuestión de cálculos, sino de atención cuidadosa a las necesidades y contextos particulares de los individuos.

Así, el dilema se vuelve inevitable: ¿debemos sacrificar la velocidad y eficiencia de las decisiones algorítmicas a cambio de una justicia más humana, necesariamente más lenta, más compleja y menos predecible? ¿O es posible construir sistemas que integren ambas dimensiones? La respuesta no es sencilla, pero la reflexión ética exige no perder de vista que la eficiencia, por más seductora que sea, no es un valor absoluto, sino un medio que debe estar siempre subordinado a la justicia y a la virtud.

Frente al entusiasmo tecnocrático por la eficiencia algorítmica, emerge una corriente crítica que advierte sobre los riesgos de legitimar decisiones automatizadas en esferas donde están en juego los derechos fundamentales. Desde una perspectiva deontológica, que enfatiza principios y deberes más allá de las consecuencias, la pregunta ética no es simplemente "¿funciona el algoritmo?", sino "¿es moralmente aceptable tratar a un ser humano como un dato procesable?". La automatización de decisiones públicas en ámbitos como el trabajo, la seguridad o el acceso a derechos sociales muestra dilemas que ninguna lógica utilitaria puede resolver del todo.

La deontología kantiana se vuelve especialmente relevante aquí. Immanuel Kant (1993) afirmaba que los seres humanos deben ser tratados siempre como fines en sí mismos, nunca como medios para un fin, y esta máxima entra en fricción directa con el diseño algorítmico, que opera necesariamente bajo criterios de clasificación, optimización y descarte. Cuando un sistema de reclutamiento automático descarta miles de candidatos por no ajustarse a un perfil previamente entrenado, tal y como ocurrió con el sistema de Amazon que discriminaba sistemáticamente a mujeres por aprender de datos históricos sesgados, el algoritmo no solo comete un error técnico: viola el principio de dignidad que Kant coloca en el centro de la ética.

Más allá de esta crítica filosófica general, las objeciones prácticas se vuelven aún

más urgentes cuando se analiza cómo los algoritmos reproducen y amplifican estructuras preexistentes de desigualdad. La filosofía política contemporánea, particularmente las teorías de justicia de John Rawls, ofrece un marco robusto para evaluar estas tensiones. En *A Theory of Justice* (1971), Rawls propone el principio de la diferencia: las desigualdades sólo son moralmente justificables si benefician a los más desfavorecidos. Aplicado a los sistemas algorítmicos, este principio implica que su diseño y aplicación deberían priorizar activamente la equidad, no simplemente la eficiencia.

Pero los hechos revelan lo contrario. Casos como el algoritmo COMPAS, utilizado en Estados Unidos para predecir la reincidencia criminal, mostraron cómo los sistemas replicaban sesgos raciales históricos, asignando mayores probabilidades de reincidencia a personas afroamericanas sin una base sólida. Lo más inquietante es que, al estar empaquetados bajo el manto de la objetividad matemática, estos sistemas operan sin que se cuestione su legitimidad moral. Como afirma Ruha Benjamin en *Race After Technology (2019)*, "el racismo no se elimina con la tecnología; se reconfigura, se codifica, se automatiza".

Este tipo de automatización, además, erosiona una de las condiciones fundamentales de la democracia: la posibilidad de rendición de cuentas. Cuando las decisiones son delegadas a cajas negras algorítmicas, los ciudadanos pierden la capacidad de apelar, entender o confrontar aquello que los afecta directamente. En este sentido, la crítica no es solamente moral o teórica, sino estructural: estamos frente a un desbalance de poder que favorece a quienes controlan los modelos, los datos y los fines de los sistemas.

Lo que está en juego es más que la equidad: es la arquitectura misma de los derechos. Un ejemplo contundente lo ofrece el sistema SyRI (System Risk Indication) implementado en Países Bajos, que cruzaba datos personales para detectar posibles fraudes en prestaciones sociales. El sistema fue finalmente declarado ilegal por un tribunal, que reconoció su violación del derecho a la privacidad y a no ser discriminado. La sentencia no sólo tumbó un algoritmo, sino que marcó un precedente sobre los límites éticos y jurídicos de la vigilancia automatizada en nombre de la eficiencia estatal.

El filósofo Byung-Chul Han ha advertido en *La Sociedad de la Transparencia (2017)* que el poder en la era digital ya no se ejerce por represión, sino por transparencia y rendimiento, convirtiendo al ciudadano en un sujeto auto expuesto, monitorizado y optimizado. En este paisaje, los algoritmos actúan como mediadores de

una racionalidad neoliberal que disuelve el espacio del juicio ético, reemplazandolo por métricas de rendimiento. Como señala también Shoshana Zuboff, la lógica del *surveillance capitalism (2019)* convierte la vida humana en un insumo para predicción y control de comportamientos, con impactos directos en la autonomía y la libertad.

Frente a este panorama, las críticas éticas no pueden quedarse en una apelación abstracta a la "equidad". Se requiere repensar los fundamentos del diseño algorítmico desde un enfoque de derechos humanos, que reconozca que la justicia no puede reducirse a eficiencia, ni la moral a precisión. Judith Butler en *Gender Trouble: Feminism and the Subversion of Identity (1990)* ha insistido en que los marcos normativos que deciden quién merece ser reconocido, protegido o incluso llorado, están cargados de sesgos culturales, y los algoritmos, por su diseño entrenado en datos históricos, replican esas mismas exclusiones.

En este sentido, la solución no pasa por "deshumanizar menos", sino por rehumanizar el proceso desde su origen. Los marcos algorítmicos deben partir de una ética que reconozca la pluralidad, la vulnerabilidad y la dignidad como ejes centrales. Es posible pensar en un diseño basado no solo en la prevención de errores, sino en el fomento activo de la justicia social.

La crítica, entonces, no es a la tecnología en sí, sino al modelo político y ético que la sostiene. La pregunta no es si los algoritmos pueden ser justos, sino bajo qué condiciones, con qué límites y al servicio de quién. Porque en última instancia, cada algoritmo que decide sobre nuestras vidas está traduciendo una visión del mundo. Y cuando esa visión se disfraza de neutralidad, lo que hace es consolidar las injusticias del presente en nombre de un futuro eficiente.

La historia de la inteligencia artificial pública no está escrita en piedra. Si los sistemas automatizados que hoy rigen procesos vitales son, en muchos casos, máquinas de reproducir injusticias, también pueden convertirse en herramientas de emancipación. Pero eso exige una transformación profunda en cómo se diseñan, implementan y supervisan. No basta con
corregir errores aislados: se trata de disputar el poder algorítmico y someterlo a formas de control democrático.

Una de las condiciones más reiteradas por los marcos éticos internacionales es la transparencia algorítmica. Sin embargo, la transparencia no puede entenderse como una apertura técnica, sino como una condición para el debate público. Como recuerda Floridi en *The ethics of algorithms: Mapping the debate*

(2016), “ninguna caja negra puede ser parte de una sociedad justa”. La explicabilidad no es solo un lujo filosófico, sino un requisito político: si no podemos entender cómo una decisión nos afecta, no podemos disputar su legitimidad. Este principio se recoge, por ejemplo, en la propuesta de la Unesco sobre la ética de la inteligencia artificial, que exige que los sistemas sean comprensibles para usuarios no expertos y que haya mecanismos de apelación claros.

Ahora bien, la transparencia no es suficiente si no va acompañada de responsabilidad institucional. Nos lo advierte Sandra Wachter también en *The ethics of algorithms: Mapping the debate (2016),* sin una estructura legal que atribuya responsabilidad en caso de daño, la transparencia corre el riesgo de convertirse en mera performatividad. Esto implica establecer líneas claras de rendición de cuentas, incluso cuando el daño resulte de sistemas automatizados. Algunas propuestas apuntan a la creación de agencias públicas independientes para la auditoría y fiscalización de la IA, con la capacidad de imponer sanciones y exigir cambios estructurales.

Un paso significativo hacia esta dirección es el *Global Index on Responsible AI 2024*, una evaluación comparativa de 138 países basada en indicadores que miden desde marcos regulatorios hasta el nivel de inclusión de actores sociales en el diseño de políticas tecnológicas. El informe muestra una paradoja central: si bien más de 80% de los países ha adoptado discursos sobre el uso responsable de la IA, muy pocos han traducido esos principios en mecanismos vinculantes. En América Latina, por ejemplo, países como Uruguay y México aparecen con avances desiguales: mientras el primero ha generado espacios de participación intersectorial, el segundo sigue careciendo de una legislación específica con dientes regulatorios.

Esta disparidad se refleja también en la escasa inclusión de comunidades vulnerables en la toma de decisiones. Si la IA reproduce desigualdades históricas, no puede regularse desde las mismas élites que han concentrado el poder tecnológico. Aquí, la noción de "justicia algorítmica participativa", propuesta por autores como Virginia Dignum en *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way (2019)*, apunta a una reforma estructural: no solo auditar los algoritmos desde afuera, sino involucrar a quienes serán afectados por ellos en su concepción misma.

Otra vía potente ha sido planteada por Christian Sandvig y su propuesta de auditorías algorítmicas críticas en *Data and discrimination: Collected essays (2014)*. A diferencia de las auditorías

corporativas, centradas en la eficiencia, este enfoque promueve una evaluación política de los impactos sociales. La auditoría no es solo un procedimiento técnico, sino una forma de contralor ciudadano, especialmente cuando se vincula con medios independientes, universidades y organizaciones sociales.

En este sentido, el trabajo de Shoshana Zuboff sobre *The age of surveillance capitalism: The fight for a human future at the new frontier of power (2029)* se vuelve crucial. Zuboff denuncia la colonización de la vida cotidiana por plataformas que convierten nuestros datos en mercancía y poder. Aplicado al ámbito público, esto supone una amenaza doble: no solo se extraen datos sin consentimiento pleno, sino que se utilizan para justificar exclusiones, monitoreos y castigos automatizados. Por eso, la regulación no puede limitarse al uso estatal de IA, sino que debe enfrentar el poder estructural de las plataformas tecnológicas privadas que moldean los algoritmos.

Por último, el trabajo de Brent Mittelstadt en *The ethics of algorithms: Mapping the debate (2016)* aporta una dimensión ética clave: la noción de "legitimidad epistémica". Si aceptamos que las máquinas producen conocimiento sobre nosotros, entonces debemos preguntarnos: ¿qué tipo de conocimiento es ese? ¿Con base en qué supuestos se produce? Mittelstadt subraya que, sin una base crítica que cuestione los datos de entrada, los algoritmos seguirán amplificando injusticias. Es decir, la ética algorítmica no puede limitarse al uso del algoritmo, sino que debe cuestionar el conocimiento que presume producir.

Lo que emerge, entonces, no es un manual de soluciones técnicas, sino una apuesta por una política tecnológica distinta, que reconozca que cada decisión automatizada es, en el fondo, una decisión sobre qué tipo de sociedad queremos construir. No se trata solo de domesticar a la máquina, sino de redistribuir el poder que la sostiene.

La historia de la técnica está plagada de promesas de emancipación que terminaron reforzando viejas cadenas. No por malicia, sino por omisión: porque cuando una herramienta se asume neutral, dejamos de cuestionar las manos que la programan y los intereses que la guían. La IA no nos ha traído el futuro: ha replicado el pasado con una eficiencia que roza lo siniestro. La ilusión de objetividad ha permitido que viejas desigualdades se maquillen como decisiones optimizadas, liberadas de ideología. Pero lo que no se nombra no desaparece; simplemente actúa desde las sombras del sistema.

En el *Global Index on Responsible AI 2024*, sólo unos pocos países alcanzan estándares aceptables en gobernanza, participación y derechos. Las voces que

más deberían contar: las de quienes padecen los algoritmos sin entenderlos ni poder apelar a ellos, siguen ausentes en la conversación. Lo que este índice muestra no es un mapa de progreso, sino una cartografía del poder: quién diseña, quién regula, quién obedece. Y lo que está en juego no es menor. No hablamos de ajustar variables o mejorar métricas, sino de disputar el sentido mismo de lo justo, lo humano, lo posible.

Quizá el mayor desafío ético no consista en domesticar a la IA, sino en no dejar que ella nos domestique a nosotros. Que no nos convenza de que el juicio humano es prescindible, que el verdadero conflicto es la ineficiencia, que la equidad puede medirse en líneas de código. La justicia, como recordó Rawls, no es una función de cálculo, sino una construcción política, histórica, frágil. Exige deliberación, escucha, conflicto y revisión constante. No se programa: se disputa.

Por eso, el debate sobre inteligencia artificial no es técnico. Es ético. Es político. Es filosófico. Porque detrás de cada decisión automatizada hay una pregunta que no puede responderse con datos: ¿qué tipo de sociedad estamos dispuestos a tolerar? ¿Y qué futuro merecemos imaginar?

Quizás aún estemos a tiempo de volver a mirar de frente el espejo que construimos. No para confirmar lo que somos, sino para corregir lo que no deberíamos seguir siendo.

Referencias

- Aristotle. (2009). *Nicomachean ethics* (W. D. Ross, Trans.). Oxford University Press. (Original work published ca. 350 B.C.E.)
- Benjamin, R. (2019)**.** *Race after technology: Abolitionist tools for the new Jim code*. Polity. Butler, J. (1990). *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.
- Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*.
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
- Foot, P. (2001). *Natural goodness*. Oxford University Press.
- Global Index on Responsible AI. (2024). *Global Index on Responsible AI 2024*. AlgorithmWatch & International Research Centre on Artificial Intelligence (IRCAI). https://www.responsible-ai.org/index/
- Han, B.-C. (2017). *La sociedad de la transparencia*. Herder Editorial.
- Kant, I. (1993). *Grounding for the metaphysics of morals* (J. W. Ellington, Trans.).

Hackett Publishing Company. (Original work published 1785)

- MacIntyre, A. (1984). *After virtue: A study in moral theory* (2nd ed.). University of Notre Dame Press.
- Mill, J. S. (1863). *Utilitarianism*. Parker, Son, and Bourn.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21. https://doi.org/10.1177/2053951716679679
- Rawls, J. (1971). *A theory of justice*. Harvard University Press.
- Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Data and discrimination: Collected essays*. Open Technology Institute.
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Derechos Humanos e Inteligencia Artificial: Análisis de cómo los sistemas de IA pueden respetar, proteger o violar los derechos humanos.

Autora: María Fernanda Vilchis Aguiñaga

La Inteligencia Artificial (IA) ha progresado aceleradamente en los últimos años, modificando de forma notable varios aspectos de nuestra vida diaria. Desde la automatización de industrias se gasta el estudio de grandes cantidades de información para tomar decisiones en diferentes campos como la educación, profesiones, entre otras. No obstante, este progreso presenta retos éticos esenciales. Específicamente la IA está provocando inquietudes respecto al respeto y la salvaguarda de los derechos humanos dado que su aplicación genera ciertos cuestionamientos acerca de la privacidad, la no discriminación y la equidad.

En la era digital, la inteligencia artificial (IA) ha comenzado a ocupar un lugar central en la toma de decisiones que afectan profundamente la vida de las personas, desde la asignación de créditos y empleos hasta la administración de justicia. Aunque se le atribuye una supuesta objetividad, la IA reproduce y amplifica los sesgos sociales contenidos en los datos con los que es entrenada. Esta automatización, si no es regulada y supervisada con criterios éticos y legales claros, puede vulnerar derechos fundamentales como la no discriminación y la igualdad ante la ley. Casos recientes documentados en el GIRAI Spanish Report 2024 muestran cómo algoritmos aplicados en salud, empleo y seguridad pública han perpetuado injusticias históricas bajo una falsa neutralidad tecnológica. Por ello, es urgente repensar el diseño, implementación y control de estas tecnologías, asegurando su transparencia, rendición de cuentas y la posibilidad de intervención humana, especialmente cuando están en juego los derechos de las personas más vulnerables.

Según el GIRAI Spanish Report 2024, la IA no es neutral. Aprende de nuestros prejuicios, los reproduce y, en muchos casos, los amplifica. ¿Cómo puede una

tecnología que no entiende de ética garantizar derechos como la igualdad o la privacidad? Al mismo tiempo, hay ejemplos en los que la IA ha ayudado a identificar injusticias o a distribuir recursos de manera más equitativa. Pero estos avances no ocurren por casualidad: dependen de quién diseña los sistemas y con qué propósito.

El incremento exponencial de esta herramienta ha expandido la habilidad de las máquinas para llevar a cabo funciones que anteriormente solo eran posibles para los humanos. Además, esta tecnología no está libre de peligros y potenciales efectos adversos. Cuando se aplica de forma irresponsable o sin una supervisión apropiada, la IA puede llegar a infringir derechos humanos esenciales, tales como el derecho a la privacidad, la libertad personal y el derecho a ser tratado de forma justa ante la ley. Este artículo analiza cómo los sistemas de inteligencia artificial pueden respetar o transgredir los derechos humanos y no busca demonizar la tecnología, sino cuestionar su uso. Porque detrás de cada algoritmo hay intereses, decisiones humanas y, sobre todo, consecuencias reales. Si la IA va a moldear nuestro futuro, es urgente decidir si lo hará para defender los derechos humanos o para socavarlos, y asegurar un uso ético y responsable de la IA.

¿De qué manera los sistemas de Inteligencia Artificial pueden asegurar el respeto a los derechos humanos y qué acciones éticas deben implementarse para evitar que la IA los infrinja como en los aspectos de la privacidad y la no discriminación?

La Inteligencia Artificial ha probado ser un recurso potente capaz de enriquecer nuestras vidas de forma notable. Sin embargo, su uso también presenta serios retos éticos. Los sistemas de inteligencia artificial, basados en algoritmos que examinan y manejan grandes datos, pueden emplearse para tomar decisiones que impactan directamente en los derechos esenciales de los individuos. Hablando sobre esto, surgen dudas acerca de ¿Cómo asegurar que la Inteligencia Artificial no viole los derechos humanos? Para seguir con el tema, este artículo se enfocará en tres áreas fundamentales donde la Inteligencia Artificial puede impactar los derechos humanos.

Cuando una máquina toma decisiones que afectan la vida de un individuo como otorgarle un crédito, evaluarlo para un empleo o clasificarlo en un sistema judicial no estamos ante un procedimiento neutral, sino ante un acto con implicaciones morales y políticas. Si estas decisiones se basan en datos sesgados, históricos o incompletos, el riesgo de reproducir y amplificar desigualdades estructurales es alarmante. Por ello, es fundamental

abordar los desafíos que la IA representa para los derechos humanos desde una perspectiva integral y proactiva, que no se limite a reaccionar ante los abusos, sino que se adelante a ellos mediante principios éticos sólidos, regulación efectiva y mecanismos de rendición de cuentas.

Este análisis se enfocará en tres dimensiones críticas donde la IA puede socavar las garantías fundamentales: el derecho a la privacidad, constantemente amenazado por sistemas de vigilancia masiva y uso indebido de datos; el derecho a la no discriminación, en peligro por algoritmos que reproducen estigmas sociales invisibilizados en los datos; y el derecho a la igualdad ante la ley, desafiado por tecnologías que sustituyen la justicia con decisiones opacas e inapelables. Reconocer estas amenazas es el primer paso hacia un desarrollo tecnológico verdaderamente humano y justo.

La Inteligencia Artificial (IA) no es solo una herramienta tecnológica, sino una fuerza social que redefine cómo vivimos, cómo nos relacionamos y cómo somos juzgados. Su avance es imparable, pero su impacto en los derechos humanos sigue siendo una incógnita que depende, no de la tecnología en sí, sino de cómo decidimos gobernarla. El GIRAI Spanish Report 2024 advierte que, sin regulaciones éticas claras, la IA puede convertirse en un mecanismo de opresión sistémica, incluso cuando se presenta como un instrumento de progreso.

El derecho a la privacidad, en el contexto de la inteligencia artificial, ha sido erosionado de forma sistemática por la lógica de recolección masiva de datos que impera tanto en el sector privado como en el público. Tal como lo advierte el Reporte GIRAI 2024, la IA ha contribuido a la consolidación de un ecosistema de vigilancia donde los datos personales se recopilan, analizan y comercializan con escasa o nula intervención consciente por parte del usuario. Este entorno, que promueve una lógica de control algorítmico, transforma la privacidad en una moneda de cambio más que en un derecho inherente.

La amenaza no radica únicamente en la existencia de tecnologías de reconocimiento facial, rastreo biométrico o geolocalización, sino en su despliegue sin mecanismos robustos de consentimiento informado. Muchas plataformas digitales y aplicaciones de uso cotidiano, bajo la fachada de ofrecer mejoras en la experiencia del usuario, ocultan prácticas intrusivas de minería de datos. El informe destaca que esta opacidad informativa vulnera el principio de autodeterminación informativa, esencial para que los individuos puedan decidir sobre el destino y uso de sus

datos personales.

Una de las formas más insidiosas en que esta vulneración ocurre es a través del consentimiento implícito: extensos y ambiguos términos y condiciones que los usuarios rara vez leen o entienden. En este sentido, el GIRAI 2024 plantea que la IA ha facilitado una "normalización del rastreo", donde aceptar la vigilancia se convierte en una condición para participar en la vida digital. Esta situación genera un círculo vicioso: para acceder a servicios esenciales como transporte, salud, empleo o educación en línea se exige entregar datos, convirtiendo el derecho a la privacidad en un lujo inalcanzable.

Desde una perspectiva ética, la situación exige una transformación profunda basada en transparencia algorítmica y justicia de datos. Las instituciones públicas y privadas deben ser obligadas a declarar de manera clara y accesible qué datos recolectan, con qué fines, y por cuánto tiempo los almacenan. Esta transparencia debe ser acompañada por mecanismos de consentimiento explícito, informado y revocable. Además, se requiere establecer límites estrictos sobre el uso de datos sensibles como origen étnico, género, orientación sexual, salud mental, afiliación política o religión que no deberían usarse para ningún tipo de segmentación, clasificación o perfilado sin justificación legal y consentimiento expreso.

Otra acción ética indispensable es la creación de organismos independientes de supervisión que auditen el cumplimiento de las normas de protección de datos, especialmente en contextos de poder asimétrico como el Estado o grandes corporaciones tecnológicas. Estos organismos deben tener la capacidad no solo de sancionar, sino también de suspender el funcionamiento de sistemas de IA que vulneren sistemáticamente la privacidad.

Además, es necesario repensar el modelo de desarrollo tecnológico dominante. En lugar de una IA centrada exclusivamente en la eficiencia y el beneficio económico, debe impulsarse una IA centrada en los derechos, donde el diseño de algoritmos priorice la protección de la dignidad humana. Esto implica aplicar el principio de privacidad por diseño, integrando salvaguardas desde las etapas iniciales de creación de sistemas, no como añadidos posteriores.

Finalmente, es crucial fomentar la educación ciudadana sobre la privacidad digital. No se puede defender un derecho que no se comprende. Por ello, la alfabetización digital debe incluir no solo habilidades técnicas, sino también conocimiento sobre derechos digitales, riesgos del uso de datos y mecanismos de protección. Así, los ciudadanos podrán

ejercer una vigilancia crítica sobre las tecnologías que usan y exigir mayores estándares éticos.

El derecho a la no discriminación es uno de los principios más vulnerados por el uso inadecuado de la inteligencia artificial. Aunque muchas veces se asume que los algoritmos son neutrales, el Informe GIRAI 2024 muestra que estos sistemas heredan y amplifican los sesgos presentes en los datos con los que se entrenan. Esto ocurre porque la IA no entiende el contexto social ni los principios éticos; simplemente reproduce patrones históricos, que muchas veces reflejan discriminación sistémica.

Los sistemas de IA no son objetivos. Aprenden de datos históricos cargados de prejuicios y los reproducen a escala industrial. La IA, por sí sola, no tiene intenciones discriminatorias; sin embargo, actúa conforme a los datos que se le entregan. Si estos datos reflejan desigualdades estructurales, las decisiones automatizadas pueden resultar injustas o excluyentes.

El reporte documenta casos donde algoritmos usados en contrataciones, préstamos bancarios o incluso en el sistema judicial han discriminado a mujeres, minorías étnicas y personas en situación de pobreza por el simple hecho de que sus datos no se ajustaban al perfil histórico del "éxito" en determinadas empresas, un perfil sesgado por décadas de exclusión sistemática. Esto se extiende también a sectores como la justicia predictiva, el otorgamiento de créditos o el acceso a seguros de salud. Lo grave no es solo que esto ocurra, sino que se esconde detrás de la falsa neutralidad de la tecnología. Para evitarlo, se necesitan auditorías independientes obligatorias que expongan los sesgos de estos sistemas, junto con sanciones reales para quienes los diseñan sin considerar el impacto social.

Uno de los ejemplos más alarmantes que presenta el informe es el caso de un algoritmo de salud en España que asignaba menos recursos a personas del pueblo gitano, no por su estado de salud, sino por su código postal, una variable correlacionada con pobreza y marginación. De manera similar, plataformas de empleo en México filtraban automáticamente a mujeres madres bajo la justificación de "menor disponibilidad laboral", ignorando por completo sus capacidades, experiencia o motivación.

Estas decisiones algorítmicas pueden parecer técnicamente justificadas desde la lógica del "perfil óptimo", pero éticamente resultan profundamente injustas. Lo más problemático es que esta discriminación se oculta bajo la fachada de la automatización, lo que dificulta la

rendición de cuentas y deja a las víctimas sin posibilidad clara de defensa.

La regulación actual es complaciente, auditar algoritmos no basta. Debe exigirse: Análisis técnicos y sociales que permitan identificar discriminaciones no intencionales en los modelos de IA. Derecho a réplica humana, cualquier decisión automatizada que afecte derechos debe poder apelarse ante una persona. Además, se propone la implementación de marcos regulatorios específicos que establezcan sanciones frente al uso discriminatorio de tecnologías automatizadas, reforzando el principio de igualdad de oportunidades. Se deben establecer leyes que prohíban expresamente el uso de tecnologías automatizadas que reproduzcan o perpetúen discriminaciones, incluyendo sanciones económicas y legales para quienes implementen estos sistemas sin una evaluación previa de impacto social. Los sistemas de IA pueden respetar los derechos humanos si se diseñan y regulan bajo un marco ético sólido. Pero si se dejan al libre albedrío de intereses comerciales o de eficiencia institucional, terminarán por reforzar y sofisticar las formas de exclusión existentes. Asegurar la no discriminación requiere voluntad política, compromiso social y una ética digital que ponga a la dignidad humana en el centro de la innovación.

El derecho a la igualdad ante la ley es uno de los más sensibles frente a la expansión de tecnologías de inteligencia artificial en los sistemas de justicia y seguridad. La aplicación de IA en procesos judiciales, policiales y administrativos representa un terreno delicado. De acuerdo al reporte, sistemas de reconocimiento facial o predicción de reincidencia criminal están siendo utilizados sin la debida supervisión, lo que puede vulnerar el derecho a un juicio justo o incluso legitimar prácticas discriminatorias por perfilamiento racial. También menciona ejemplos donde personas han sido injustamente perjudicadas por errores algorítmicos, sin posibilidad de apelación porque "la máquina lo decidió". La automatización de decisiones en contextos legales no elimina la posibilidad de error, sino que la desplaza a una caja negra algorítmica. Si los ciudadanos no pueden apelar, entender o cuestionar una decisión tomada por un sistema de IA, se rompe el principio de rendición de cuentas. El problema se agrava cuando estas decisiones automatizadas no permiten un análisis transparente ni ofrecen al ciudadano una vía efectiva de apelación. Esto vulnera gravemente el principio de igualdad ante la ley, que exige que toda persona sea juzgada en condiciones equitativas y con acceso a un proceso justo.

La igualdad ante la ley exige transparencia, nadie puede ser juzgado por un sistema cuyo funcionamiento se desconoce. Por eso, es urgente exigir que

toda IA usada en ámbitos legales sea explicable, auditada y, sobre todo, toda decisión que afecte gravemente los derechos de una persona debe contar con la posibilidad de intervención o revisión humana. Las personas deben tener derecho a entender en términos razonables cómo y por qué una IA ha tomado cierta decisión sobre ellos. No se debe justificar el uso de IA con fines de seguridad sin evaluar previamente si es estrictamente necesario y si existen medios menos invasivos. Se debe proteger el equilibrio entre seguridad y libertades individuales.

Se resalta también la necesidad de criterios de proporcionalidad y necesidad para el uso de tecnologías de vigilancia, de modo que se respeten los principios democráticos en contextos sensibles como la seguridad pública. La IA puede ser una herramienta valiosa en el ámbito legal si se usa con responsabilidad, supervisión y bajo un marco ético robusto. Pero si se convierte en una fuente incuestionable de autoridad, sin espacio para el escrutinio ciudadano o el control democrático, terminará por socavar los principios de justicia e igualdad ante la ley. La tecnología debe estar al servicio de la justicia, no por encima de ella.

La inteligencia artificial puede ser una aliada de la justicia, pero sólo si se integra dentro de un marco legal robusto que garantice transparencia, rendición de cuentas y revisión humana efectiva. Automatizar sin ética es delegar la justicia a sistemas incapaces de comprender lo humano. En cambio, una IA al servicio de los derechos, diseñada con participación social y bajo principios democráticos, puede ayudar a reducir sesgos humanos, pero solo si primero se reconocen y corrigen los de su propia arquitectura.

La paradoja es clara, la misma tecnología que puede optimizar recursos médicos o predecir crisis climáticas también puede vigilar, excluir y juzgar sin piedad. La promesa de la IA como herramienta de mejora social debe ir acompañada de un compromiso firme con los derechos humanos. El reporte GIRAI 2024 evidencia que aún queda un largo camino por recorrer para alinear el desarrollo tecnológico con los valores éticos que sustentan una sociedad democrática. No se trata solo de mejorar los algoritmos, sino de repensar los marcos normativos, los procesos de diseño y los fines para los que usamos estas tecnologías.

El verdadero progreso no se medirá únicamente por la eficiencia o la innovación, sino por la capacidad de construir sistemas que respeten la dignidad humana, promuevan la justicia social y fortalezcan la equidad.

Asegurar que la Inteligencia Artificial respete los derechos humanos exige una acción coordinada entre reguladores, empresas, desarrolladores y ciudadanos.

Necesitamos sistemas diseñados desde una ética de la responsabilidad, con mecanismos de control que anticipen y corrijan los impactos negativos, y con una profunda convicción de que la tecnología debe servir a las personas, y no al revés. La IA debe ser diseñada y aplicada bajo un principio fundamental: la dignidad humana no es negociable. Esto implica reconocer que derechos como la privacidad, la igualdad ante la ley y la no discriminación no deben subordinarse a intereses económicos, comerciales o de eficiencia.

Una propuesta ética para el desarrollo y aplicación de la inteligencia artificial debe partir de un marco de gobernanza integral que articule valores humanos fundamentales con la innovación tecnológica. En este sentido, es imprescindible que los sistemas de IA se diseñen desde una lógica de codiseño con impacto social, lo cual significa incluir, desde las fases iniciales del desarrollo, a comunidades afectadas y grupos vulnerables, asegurando que sus necesidades y riesgos específicos sean considerados. Esto contribuye a evitar que la IA perpetúe o amplifique desigualdades ya existentes. Además, debe establecerse la rendición de cuentas obligatoria como principio estructural: cada decisión tomada por una IA debe tener un responsable humano o institucional claramente identificable, lo que permite exigir justicia cuando se vulneran derechos. Esta responsabilidad no puede delegarse completamente a la máquina, pues hacerlo diluye la noción misma de justicia. Paralelamente, se requiere una educación digital en derechos humanos que forme tanto a desarrolladores como a usuarios en ética, justicia social y regulación, de modo que el conocimiento técnico se complementa con una conciencia crítica. Finalmente, es vital instaurar mecanismos de supervisión continua e independiente, mediante organismos autónomos y comités interdisciplinarios capaces de auditar, detener o reformular tecnologías que atenten contra los principios fundamentales de equidad, dignidad y privacidad. Solo a través de este enfoque holístico y proactivo se puede garantizar que la inteligencia artificial avance sin comprometer los derechos humanos que sostienen nuestras sociedades democráticas.

Referencias

- Access Now. (2021). Derechos humanos en la era de la inteligencia artificial. Alianza Global Jus Semper. https://jussemper.org/Inicio/Recursos/Democracia%20Mejores%20Practicas/Resources/Acce ssNow-AlyDerechosHumanos.pdf

- GIRAI. (2024). Informe GIRAI Spanish Report 2024. https://girai-spanish-report 2024.tiiny.site/
- Grigore, A. E. (2022). Derechos humanos e inteligencia artificial. https://revistascientificas.us.es/index.php/ies/article/view/19991

Humanidad reducida a datos

Autora: Roxana Michelle Rosas Vega

¿Es moralmente aceptable que la automatización de procesos transforme a los seres humanos en meros datos cuantificables, reduciéndose a patrones predecibles y negando su singularidad, autonomía y complejidad moral para la toma de decisiones? Esta pregunta se vuelve cada vez más urgente a medida que la inteligencia artificial deja de ser una herramienta pasiva para convertirse en un agente activo dentro de nuestras estructuras sociales. Lo que comenzó como un sistema programado para seguir instrucciones básicas, hoy se infiltra en las estructuras donde se decide el rumbo de empresas y gobiernos. La inteligencia artificial ya no es solo una herramienta: es un agente que participa, influye y, en ocasiones, redefine la toma de decisiones. Con el incremento del *Big Data* y el *machine learning*, la capacidad de habilidades y tareas que son transferidas a la IA incrementa de una manera rápida, prediciendo comportamientos, evaluando riesgos y hasta llegando a reemplazar deliberaciones humanas. En poco tiempo pasamos de usar una herramienta que simplemente ejecuta instrucciones, a convivir con un sistema que decide quién accede a un tratamiento médico, quién recibe libertad condicional o quién consigue un empleo. Este veloz desplazamiento de lo humano a lo algorítmico da lugar a distintas opiniones y posturas. ¿Es éticamente correcto confiar decisiones de alto impacto a algoritmos que, aunque garantizan mayor imparcialidad y consistencia, carecen de conciencia moral y empatía, o es más adecuado que los seres humanos, pese a sus limitaciones, sigan asumiendo esta responsabilidad para asegurar que las decisiones reflejen valores humanos esenciales?

La promesa de la inteligencia artificial ha sido bastante tentadora, ya que optimiza procesos de manera impresionante, y que en apariencia promete reducir los sesgos humanos. No obstante, en su afán de predecir y clasificar, la IA convierte al ser humano en un simple conjunto de datos cuantificables. Reduce la experiencia individual a patrones y, al hacerlo, se pierden valores humanos fundamentales como la empatía, la dignidad, etc. Lo que en un principio parecía un avance técnico se revela también como una transformación

moral. Por lo que esta problemática va más allá de la eficacia de los algoritmos, también se trata de preguntarnos qué clase de humanidad estamos creando cuando permitimos que tomen decisiones por nosotros y quiénes son los que están diseñando estos sistemas y que intenciones o dilemas de poder están implícitos en los algoritmos.

Un argumento muy común a favor de la automatización es el concepto de que las máquinas tienen la capacidad de ser más "imparciales" que los seres humanos. Al no contar con emociones y prejuicios inconscientes, la Inteligencia Artificial promete tomar decisiones más equitativas y consistentes. (Cowgill, Dell'Acqua, & Deng, 2021). En una investigación realizada por la Harvard Business School, se descubrió que numerosos directivos optan por decisiones adoptadas por Inteligencia Artificial en ámbitos financieros, justamente por su aparente imparcialidad. No obstante, desde una perspectiva deontológica, no es suficiente con valorar las consecuencias. (De Cremer & Kasparov, 2021). Según la ética de Immanuel Kant, el valor moral de una acción no radica en sus consecuencias, sino en si la acción misma respeta el deber y puede convertirse en una ley universal. En su Fundamentación de la metafísica de las costumbres (Kant, 1785/2009), plantea que los seres humanos deben ser tratados siempre como fines en sí mismos y nunca meramente como medios para un fin. Esto implica reconocer la dignidad y autonomía moral de cada persona, sin reducirla a una función instrumental.

Considerar a una persona como un dato, aunque pueda conducir a un resultado eficiente, es convertirla en un medio para lograr un objetivo. Esta lógica entra en tensión directa con el imperativo categórico kantiano, que exige actuar solo según máximas que puedan convertirse en leyes universales. Además, clasificar a los individuos en función de patrones algorítmicos —sin que estos participen en el proceso de deliberación— establece una relación asimétrica entre quienes diseñan los sistemas y quienes son evaluados por ellos (Eubanks, 2018). El sujeto ya no tiene agencia ni posibilidad de apelar: es reducido a una categoría estadística. Desde esta perspectiva, la lógica algorítmica impone un modelo de poder unidireccional, en el que el algoritmo —y por extensión sus creadores— se erige como dominador, mientras que el evaluado queda despojado de su individualidad y voz.

Un caso notable que ejemplifica esto es el sistema COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), que se utiliza en Estados Unidos y se basa en un algoritmo diseñado para predecir la reincidencia de

criminales. Sin embargo, un análisis realizado por ProPublica puso de manifiesto que este sistema tenía una tendencia notable a clasificar erróneamente a personas negras como de alto riesgo, al mismo tiempo que subestima la probabilidad de reincidencia entre las personas blancas. Al igual que el caso de Amazon en 2014, en donde se creó un algoritmo con el fin de optimizar el proceso de elección de candidatos para empleos en donde analizaba currículums y seleccionar a los candidatos que se alineaban con las necesidades. Sin embargo, con el tiempo se reveló que, de manera sistemática, penaliza a las mujeres. Debido a que el algoritmo se entrenó con datos de contrataciones anteriores, donde predominaban los hombres, lo que llevó al sistema a concluir que los currículums masculinos eran “más deseables”. Este sesgo no fue intencionado, pero generó efectos reales y discriminatorios. A pesar de que el algoritmo era técnicamente eficiente, violaba principios fundamentales de justicia. Desde una perspectiva deontológica, dicha práctica carece de justificación, ya que discriminar por género no puede ser aceptado como una ley universal válida. Asimismo, desde el enfoque utilitarista, aunque pudiera parecer eficaz al "agilizar" los procesos de selección, y que beneficiaría a la mayoría, entonces por lo tanto estaría justificado, lo cual no fue el caso.

Otro punto importante es que estos algoritmos pueden procesar grandes cantidades de información en segundos, algo imposible para los humanos. Esto permite, por ejemplo, detectar fraudes bancarios de manera inmediata o realizar diagnósticos médicos precisos. En el ámbito de la salud, sistemas como IBM Watson han sido utilizados para analizar historiales clínicos y proponer tratamientos personalizados, ayudando a los médicos a tomar decisiones mejor informadas en menos tiempo (Topol, 2019; IBM, 2020). Desde una perspectiva utilitarista, el uso de inteligencia artificial puede considerarse moralmente válido cuando mejora los resultados para la mayoría, ya que este enfoque ético sostiene que la moralidad de una acción se mide por su capacidad para maximizar el bienestar general, o bien, minimizando el mal. (Mill, 1863/2001). Además, muchas decisiones humanas están cargadas de emociones, favoritismos o cansancio, lo que puede derivar en injusticias (Kahneman, 2011). En ese sentido, la IA representa una oportunidad para elevar el estándar de equidad en procesos donde el error humano puede tener consecuencias graves.

Sin embargo, desde otra perspectiva, como mencioné anteriormente, estos algoritmos son diseñados por humanos y se basan en datos históricos que, en general, perpetúan desigualdades

globales existentes de manera sistemática. Muchos sistemas de IA implementados en países del Sur Global no consideran sus realidades sociales, económicas o culturales. Según el Global Index on Responsible AI 2024, los datos suelen originarse en centros tecnológicos del Norte Global, generando una dinámica conocida como "colonialismo algorítmico", en la que poblaciones vulnerables son objeto de recolección y procesamiento de datos sin participación ni beneficio real (Dubber, Pasquale, & Das, 2020). Este fenómeno implica la pérdida de control sobre los datos y las decisiones derivadas de ellos por parte de los países menos desarrollados, reproduciendo una forma de dominación tecnológica. Además, el *Handbook of Artificial Intelligence Ethics* señala que esta situación no solo afecta la justicia distributiva, sino también la autonomía de las comunidades, al limitar su capacidad para influir en sistemas que impactan directamente en sus vidas (Dubber, Pasquale, & Das, 2020).

Además, esta automatización dificulta el derecho de los ciudadanos a saber por qué se ha tomado una decisión sobre ellos (*accountability*). En muchos casos, ni siquiera los diseñadores de estos algoritmos comprenden del todo cómo se llegó a esa conclusión, especialmente en los sistemas que usan *machine learning*. Esta opacidad debilita principios fundamentales de responsabilidad, ya que si nadie puede explicar el resultado, tampoco se puede responsabilizar a alguien por los errores o injusticias cometidas. Asimismo, el *Handbook of Artificial Intelligence Ethics* subraya la importancia del concepto de *agency*, que se refiere a la capacidad de los individuos para ejercer control y participar activamente en las decisiones que afectan sus vidas. Cuando las decisiones automatizadas carecen de transparencia y no permiten intervención o contestación, se reduce la *agency* de las personas, limitando su autonomía y cuestionando la justicia del proceso. Este doble déficit, tanto en *accountability* como en *agency*, representa un desafío ético fundamental para la implementación responsable de la IA (Dubber, Pasquale, & Das, 2020). Un ejemplo de esto es el caso del sistema de reconocimiento facial de la policía de Detroit, Estados Unidos, en donde Robert Williams, un hombre afroamericano, fue arrestado injustamente en 2020 porque un algoritmo lo identificó erróneamente como sospechoso de robo. Las cámaras de vigilancia captaron a un hombre robando relojes, y el sistema automatizado de reconocimiento facial señaló que se parecía a Williams. A lo que la policía actuó directamente, sin una investigación previa y tomando como 100% cierto este análisis del algoritmo. (Gross, 2025). Después de pasar más de 24 horas detenido injustamente, Robert fue liberado cuando se demostró que el

algoritmo se había equivocado.

Justamente en este caso se ve reflejado lo que menciono de que no hay rendición de cuentas en este tipo de errores, las autoridades afirmaron que “el sistema no era perfecto” y que “los algoritmos a veces se equivocan”, mientras que la empresa detrás del software dijo que su herramienta era solo un “apoyo” y no debía usarse como única fuente de decisión. Así, la culpa quedó entre la tecnología y las personas que la usaron, pero sin consecuencias concretas para ninguno de los responsables directos. Desde una perspectiva aristotélica, este vacío ético es aún más problemático. Para Aristóteles, la virtud —especialmente la *phronesis* o prudencia— es esencial para actuar con justicia, ya que implica deliberación consciente, comprensión del contexto y responsabilidad moral (Aristóteles, Ética a Nicómaco). Los algoritmos, al carecer de conciencia y fines propios, no pueden actuar virtuosamente ni ser responsables. Por eso, cuando no se puede explicar cómo se llegó a una decisión, no solo se debilita la transparencia, sino que también se pierde la posibilidad de evaluar si esa acción fue éticamente correcta. Esta incapacidad de los sistemas automatizados para encarnar virtudes humanas subraya la necesidad de que las decisiones críticas no se deleguen totalmente a entidades no humanas. (Aristóteles, 2004).

Otro punto importante es la deshumanización y reducción a datos. La deshumanización se refiere al proceso de quitarle a una persona su identidad individual, convirtiéndola en un simple conjunto de datos o números. Como lo mencioné anteriormente, en un mundo donde las decisiones son cada vez más tomadas por inteligencia artificial, esta deshumanización se refleja en cómo los algoritmos tratan a las personas como simples estadísticas, ignorando su contexto, emociones o experiencias únicas. Byung-Chul Han en No-cosas señala que la digitalización y la cuantificación constantes transforman a los individuos en objetos de control y vigilancia, erosionando la experiencia subjetiva y la singularidad humana al reducir todo a datos manejables y explotables (Han, 2021). Un ejemplo claro de esta problemática es Clearview AI, una empresa que ha desarrollado un sistema de reconocimiento facial utilizado por fuerzas policiales y entidades gubernamentales. Este sistema recopila imágenes de personas desde plataformas públicas como Facebook, Instagram y otras redes sociales, sin el consentimiento de los usuarios, y las utiliza para identificar individuos en cualquier lugar con solo una foto (Angwin, Larson, Mattu, & Kirchner, 2020). Su uso se ha extendido globalmente, generando

profundas preocupaciones sobre la privacidad y la autonomía individual. Desde la perspectiva de la teoría de la justicia, este tipo de tecnología no respeta el principio de justicia como equidad, ya que el acceso a la privacidad y a la autonomía individual no es equitativo entre los ciudadanos. Una sociedad justa debe asegurarse de que las estructuras sociales y políticas beneficien en mayor medida a los más desfavorecidos. En el caso de Clearview AI, el control sobre los datos personales está en manos de empresas tecnológicas y gobiernos, dejando a los individuos vulnerables sin mecanismos efectivos para controlar o proteger el uso de su información. Esta situación es aún más injusta para las comunidades vulnerables, cuyas informaciones pueden ser usadas en su contra sin posibilidad de consentimiento o protesta.

Tomando en cuenta la dualidad de esta problemática y los argumentos que respaldan a cada una de ellas, podría responder que la idea suena útil y hasta un tanto utópica; los algoritmos pueden ser más rápidos, más consistentes, y menos emocionales que los humanos. Pero ahí está justamente el problema: ¿podemos llamar "justa" a una decisión que no tiene en cuenta las emociones, la historia personal o el contexto de una vida humana?, viéndolo así no es ético dejar completamente en manos de la IA decisiones de alto impacto humano. La razón es que estas decisiones no solo necesitan datos, sino comprensión, empatía y una valoración de lo que significa ser humano. Los algoritmos no sienten. No se preocupan por las consecuencias emocionales de una decisión. Solo optimizan para un resultado que alguien programó con criterios específicos, que muchas veces responden a intereses empresariales o institucionales, no a valores humanos. Claramente esta postura tiene límites por ejemplo, no estoy diciendo que la IA no debe usarse en decisiones importantes, sino que no debe hacerlo sola, sin supervisión o sin responsabilidad humana. Considero que la IA puede ser una herramienta poderosa si se usa con cuidado, para apoyar decisiones humanas, no para reemplazarlas. Limitar el uso de la IA en estas decisiones podría significar procesos más lentos o costosos. También podríamos desaprovechar algunas ventajas como la eliminación de sesgos personales (racismo, clasismo, etc.) que algunos humanos sí tienen y que los algoritmos, si están bien diseñados, podrían evitar. También existe el riesgo de que, si limitamos demasiado su uso, algunos países o empresas se queden atrás tecnológicamente, lo que puede tener consecuencias económicas.

Pero los riesgos de delegar decisiones clave a la IA son enormes. Si dejamos que los algoritmos decidan por nosotros,

corremos el peligro de deshumanizar procesos fundamentales. Un algoritmo no puede detenerse a considerar circunstancias especiales o contextos personales que justificarían excepciones; simplemente sigue reglas preestablecidas basadas en datos, sin cuestionar si esos datos son justos o completos. Además, cuando estos sistemas fallan, la responsabilidad se diluye: ¿la culpa recae en el programador, en la empresa, o en el propio algoritmo? Esta opacidad afecta directamente el principio de *accountability* (rendición de cuentas), pues sin transparencia ni explicabilidad, es imposible identificar a los responsables ni corregir errores. Por otro lado, está la cuestión del *agency* (agencia), o la capacidad de los individuos para participar activamente en las decisiones que les afectan. En la mayoría de los casos, las personas quedan reducidas a sujetos pasivos frente a sistemas opacos que controlan aspectos centrales de sus vidas. El poder se concentra en manos de unas pocas grandes empresas tecnológicas que desarrollan y controlan estas IA, imponiendo reglas sin participación democrática ni supervisión ciudadana. Esto configura una nueva forma de dominación algorítmica, en la que quienes toman las decisiones esenciales sobre nosotros no rinden cuentas y limitan nuestra autonomía.

Desde mi perspectiva personal al escribir este ensayo recordé una parte de un libro que estoy leyendo que se llama Sapiens de Yuval Noah Harari, en donde menciona que aparentemente los neandertales cuidaban de sus enfermos y débiles, esto lo saben porque en restos óseos se observó que tenían impedimentos físicos graves, pero que aún así vivieron muchos más años, esto significa que se apoyaban mutuamente, y esto claro que no representaba "eficiencia" o "optimización" para el grupo, al contrario, sin embargo lo hacían porque sentían empatía por el otro y considero que es algo que ningún algoritmo puede replicar. Creo que si es algo que llevamos poniéndolo en práctica por miles de años, entonces es algo inherente a nuestra existencia, y no debería de ser reemplazado por un algoritmo que no es capaz de replicar esta esencia humana. Podemos programar a una inteligencia artificial para que reconozca emociones, patrones de comportamiento o necesidades médicas. Pero eso no es lo mismo que sentir empatía. La empatía no es una fórmula ni una base de datos. Es una conexión profundamente humana que surge del reconocimiento del otro como igual, como sujeto con dignidad y valor intrínseco. Por eso, cuando trasladamos decisiones morales o sociales a sistemas automatizados, corremos el riesgo de perder una parte fundamental de lo que somos.

También al escribir la parte de la deshumanización recordé que en ese mismo libro Yuval también menciona la "teoría de peugeot" en donde básicamente explica cómo en si la empresa no existe, ya que si un juez ordena la disolución de la compañía, seguiría la fábrica y sus trabajadores, pero la empresa como tal ya no existiría, también menciona el hecho de que si corren a todos los trabajadores, estos pueden ser reemplazados y seguiría siendo Peugeot, y de ahí desarrolla la historia del cómo esta "ficción legal" pertenece a las "compañías de responsabilidad limitada" y esto se remonta al siglo XIII en donde a las personas les asustaba iniciar emprendimientos, por el miedo a asumir responsabilidad en riesgos económicos y que por eso la gente comenzó a imaginar colectivamente la existencia de estas compañías, lo que las individualizada de su responsabilidad, o sea ya no te debía la persona en concreto, te debía la empresa, y justo creo que se ve reflejado en este dilema, ya que como lo mencioné, cuando una IA se equivoca, se dice que “el algoritmo falló”, como si fuera una entidad neutral, sin relación con los humanos que la crearon y diseñaron.

Pero lo cierto es que siempre hay personas detrás: desarrolladores, corporaciones, gobiernos. (Harari, 2015). Y muchas veces, esa “neutralidad” se convierte en una excusa para escudarse. Se usa como una barrera para evitar asumir responsabilidades éticas. Al igual que en el caso de las empresas, construimos una ficción que oculta los intereses humanos y las estructuras de poder reales que operan detrás. Por lo que creo que no es muy racional pensar que automatizar a la IA para la toma de decisiones elimina ese "sesgo humano", yo diría que se ha demostrado en diversos casos que más bien lo petrifica, lo reproduce, ya que simplemente este algoritmo es un reflejo de nosotros como humanidad, ya que estos sistemas aprenden de nuestros datos, de nuestras decisiones pasadas. No vienen de otro planeta ni nacen puros. Son un reflejo de nosotros como sociedad. Y si no somos conscientes de esto, podemos terminar reproduciendo nuestras peores desigualdades bajo la ilusión de objetividad.

En definitiva, creo que automatizar la toma de decisiones no solo plantea un riesgo técnico o legal. Es, sobre todo, un riesgo moral. Porque nos obliga a cuestionarnos qué tipo de humanidad queremos construir. En el fondo, la cuestión no es simplemente qué puede hacer la inteligencia artificial, sino qué deseamos que haga en el tejido de nuestra existencia.

No se trata solo de optimizar procesos o maximizar la eficiencia, sino de

reflexionar sobre qué tipo de futuro queremos forjar: ¿uno donde la tecnología sea un espejo frío que refleja solo cálculos y datos, o uno en el que sea un instrumento para expandir nuestra humanidad? La verdadera revolución no está en el poder computacional, sino en la voluntad ética que pongamos para guiar ese poder. La tecnología, por sí sola, no posee intención ni conciencia; es un ente vacío que se llena con los valores que le imprimimos. En este sentido, el futuro que imagino es aquel donde no permitamos que la búsqueda de la optimización borre nuestra capacidad de empatía y cuidado —esas facultades que nos han sostenido milenios y que nos definen como seres conscientes y morales. Porque más allá del algoritmo y la máquina, lo que permanece irreemplazable es nuestra capacidad para reconocer al otro no como un dato, sino como un ser con dignidad, vulnerabilidad y valor intrínseco. Solo así podremos construir un mañana verdaderamente humano, no una mera simulación de vida sin alma. Automatizar decisiones críticas sin un marco ético sólido y sin participación humana consciente es renunciar a nuestra humanidad y perpetuar desigualdades bajo la apariencia de objetividad. La tecnología debe servir para ampliar nuestra capacidad de justicia y empatía, no para suplantarlas. Solo así podremos construir una sociedad donde la innovación y la dignidad humana coexistan sin contradicciones.

Referencias

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Aristóteles. (2004). Ética a Nicómaco (M. García Gual, Trad.). Alianza Editorial. (Obra original escrita ca. 350 a.C.)
- Bo Cowgill & Dell'Acqua, F. (2020). Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. arXiv. https://arxiv.org/abs/2012.02394
- Clearview AI's facial recognition technology designed for surveillance of marginalized groups, report reveals - Business & Human Rights Resource Centre. (s. f.). Business & Human Rights Resource Centre. https://www.business-humanrights.org/es/%C3%BAltimas-noticias/clearview-ais-facial-recognition-technology-designed-for-surveillance-of-marginalized-groups-report-reveals/
- Dastin, J. (2018, October 10). Amazon scrapped 'sexist AI' recruiting tool. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
- De Cremer, D., & Kasparov, G. (2021). The ethical AI paradox: How valuing consequences

leads to unethical behavior. Harvard Business School Working Paper No. 21-020. https://www.hbs.edu/ris/Publication%20Files/21-020_6fc447e4-719c-4e1c-bf1b-f8c8e0e20f9d.pdf

- Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press. https://automatinginequality.org/
- Global Index on Responsible AI. (s. f.). https://www.global-index.ai/Region-South-and-Central-America
- Gross, P. (2025, 4 de febrero). Facial recognition in policing is getting state-by-state guardrails. Stateline. https://stateline.org/2025/02/04/facial-recognition-in-policing-is-getting-state-by-state-guardr ails/
- Han, B.-C. (2021). No-cosas. https://catedradatos.com.ar/media/No-cosas-Byung-Chul-Han.pdf
- Harari, Y. N. (2015). Sapiens: De animales a dioses: Una breve historia de la humanidad. Debate.
- Harari, Y. N. (2015). Sapiens: A brief history of humankind. Harper.
- Hill, K. (2020, June 24). Wrongfully accused by an algorithm. The New York Times. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html
- Kant, I. (2009). Fundamentación de la metafísica de las costumbres (M. García Morente, Trad.). Espasa-Calpe. (Obra original publicada en 1785). https://www.ebooksgratis.mx/ebooks/filosofia/Fundamentacion%20de%20la%20Metafisica% 20de%20las%20Costumbres%20-%20Immanuel%20Kant.pdf
- Research, H. A. D. A. (2025, 15 enero). The Future of Decision-Making: How Generative AI Transforms Innovation Evaluation. Digital Data Design Institute At Harvard. https://d3.harvard.edu/the-future-of-decision-making-how-generative-ai-transforms-innovatio n-evaluation/
- Reuters. (2024, 3 de septiembre). Clearview AI fined by Dutch agency for facial recognition database. Reuters. https://www.reuters.com/technology/artificial-intelligence/clearview-ai-fined-by-dutch-agenc y-facial-recognition-database-2024-09-03/
- Rosen, A. (2020, January 27). Un hombre inocente fue arrestado debido a un error del sistema de reconocimiento facial. El País. https://elpais.com/tecnologia/2020/01/27/actualidad/1580149292_175377.html
- Ross, C., & Swetlitz, I. (2023, 31 julio). IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. STAT. https://www.statnews.com/2017/09/05/watson-ibm-cancer/

La encrucijada de la inteligencia artificial: innovación, ética y gobernanza global

Autor: Valentin Jouffret

La inteligencia artificial (IA) se ha convertido, en pocos años, en una palanca clave de transformación económica, social y política. Desde la medicina personalizada hasta la gestión energética, pasando por la optimización de los sistemas educativos o de transporte, sus aplicaciones parecen infinitas. Esta versatilidad tecnológica ha intensificado la carrera global por la innovación, en la que los países más avanzados compiten ferozmente por desarrollar y adoptar soluciones basadas en IA con mayor rapidez. En este escenario, la innovación se ha convertido en un vector de competitividad estratégica, capaz de redefinir la posición de una nación en el tablero internacional. No obstante, esta dinámica acelerada de innovación y búsqueda de ventaja competitiva plantea una serie de preocupaciones éticas fundamentales: la transparencia de los algoritmos, el respeto a los derechos humanos, el control democrático o incluso la alineación de los objetivos de la IA con los valores humanos. Así, el reto actual no consiste únicamente en innovar, sino en hacerlo de manera que preserve la dignidad, la equidad y los derechos fundamentales en un entorno cada vez más impulsado por tecnologías autónomas.

De este modo, surge una profunda tensión entre dos imperativos: por un lado, la presión por innovar rápidamente para no quedarse atrás en el escenario internacional; por otro, la necesidad de garantizar un desarrollo responsable, respetuoso de los principios éticos universales. Este dilema plantea una cuestión central: **¿deberían los países acelerar el desarrollo de la IA para mantenerse competitivos, aunque ello implique comprometer ciertos principios éticos y derechos humanos, o deberían, por el contrario, priorizar la implementación de marcos éticos sólidos, aunque eso retrase su avance tecnológico?**

Este ensayo propone analizar en profundidad esta problemática. En primer lugar, examinaremos los argumentos a favor de un desarrollo acelerado, motivado por cuestiones económicas, políticas y científicas. A

continuación, abordaremos los riesgos asociados a dicha estrategia, tanto desde el punto de vista moral como geopolítico. Finalmente, la tercera parte propondrá una vía intermedia, basada en una gobernanza internacional cooperativa, una regulación ágil y la integración de principios éticos desde el diseño de los sistemas de IA.

El desarrollo rápido de la inteligencia artificial suele presentarse como una necesidad estratégica. En un mundo globalizado y altamente competitivo, los Estados enfrentan una presión constante por innovar, especialmente en los campos tecnológicos más prometedores. La IA, por su capacidad de transformar profundamente las economías, los sistemas militares de defensa, la investigación científica y las políticas públicas, aparece como una palanca esencial para mantener o adquirir una posición de liderazgo. Por ejemplo, el uso de algoritmos de IA en la predicción y gestión de pandemias ha permitido a ciertos países, como China, anticipar respuestas sanitarias más eficaces, consolidando así su prestigio científico y su capacidad de influencia internacional.

En un contexto de creciente rivalidad entre potencias tecnológicas, especialmente entre Estados Unidos, China y la Unión Europea, un desarrollo acelerado de la IA puede verse como una condición *sine qua non* para evitar la dependencia estratégica. Dominar la IA puede entenderse en un doble sentido: por un lado, implica el desarrollo de esta tecnología a través de empresas clave que operan bajo los intereses económicos y geopolíticos de ciertos países; por otro, supone una comprensión técnica avanzada que permita diseñar, controlar e implementar sistemas inteligentes de forma autónoma. Así, dominar la IA no solo permite garantizar la autonomía en sectores críticos (salud, ciberseguridad, finanzas), sino también imponer sus propios estándares y valores. La ventaja del first mover, frecuentemente citada en la literatura estratégica, consiste en establecer normas de facto antes de que pueda surgir un consenso ético o regulador.

Como lo señala Nick Bostrom en *Superintelligence*, la primera potencia en alcanzar un nivel de inteligencia artificial general podría adquirir una ventaja tecnológica tan decisiva que se volvería inalcanzable para sus competidores. Esta perspectiva alimenta una especie de *carrera armamentista algorítmica*, en la que la rapidez de ejecución prevalece sobre la cautela ética.

Los defensores del desarrollo acelerado también destacan los beneficios tangibles que puede aportar la IA en áreas esenciales para el desarrollo del bienestar de la sociedad. En el ámbito de

la salud, permite diagnósticos más tempranos, tratamientos personalizados y una mejor asignación de recursos. En el medio ambiente, optimiza la gestión de energías renovables, mejora los modelos climáticos y facilita la transición ecológica. En educación, las tecnologías de aprendizaje adaptativo pueden reducir las desigualdades de acceso al conocimiento.

No invertir rápidamente en estas tecnologías equivaldría, bajo esta lógica, a privar a millones de personas de herramientas capaces de mejorar concretamente su calidad de vida. En este sentido, la lentitud ética en la implementación de regulaciones adecuadas y fomento a la innovación, tiene repercusiones éticas sobre la calidad de la vida de las personas.

Especialistas como Luciano Floridi, profesor de ética de la información en la Universidad de Oxford, defienden un enfoque pragmático según el cual la ética debe acompañar la tecnología sin necesariamente precederla. En sus trabajos sobre la *Ética de la infosfer*a, Floridi insiste en que la regulación no puede ser fija ni universal desde el principio, sino que debe surgir en interacción constante con los usos reales de la tecnología.

Este enfoque implica que los marcos normativos deben construirse de manera progresiva, mediante un proceso de aprendizaje social, experimentación y corrección continua. Se habla aquí de "*regulación iterativa*" o "*por diseño*", en la que los principios éticos se integran en los ciclos de desarrollo técnico, sin dejar de ser adaptables a nuevas realidades.

Además, la propia innovación puede favorecer la ética. Una IA más potente puede ser programada para detectar y corregir sesgos, hacer sus decisiones más comprensibles o incluso integrar restricciones morales en sus objetivos (como proponen los principios de la IA explicable, promovidos por la Comisión Europea). En este sentido, acelerar el desarrollo no significa ignorar la ética, sino explorar caminos tecnológicos de autorregulación algorítmica.

Floridi incluso afirma que debemos dejar de pensar en la IA como una herramienta neutra, y reconocer que toda tecnología transforma la estructura moral de la acción humana: se trata, entonces, de desarrollar la IA con ética, y no contra ella.

Otro argumento a favor del desarrollo acelerado de la IA es la posibilidad de reducir ciertas desigualdades estructurales, especialmente en el Sur global. La IA, por su capacidad de ofrecer soluciones escalables y de bajo coste, puede ayudar a resolver problemas sistémicos: diagnóstico médico a

distancia, análisis predictivo de cosechas, traducción automática para lenguas poco representadas o detección temprana de catástrofes naturales. Con acceso a estas tecnologías, comunidades históricamente marginadas podrían saltar etapas del desarrollo (lo que a veces se llama *leapfrogging tecnológico*), como ocurrió con la telefonía móvil en África.

Expertos como Kai-Fu Lee, ex presidente de Google China, destacan que ralentizar el progreso en los países más avanzados podría retrasar el acceso a estas innovaciones en las regiones que más las necesitan. Para él, la IA podría ser un motor de crecimiento inclusivo, siempre que se desarrolle de forma masiva y abierta.[3]

Sin embargo, como lo subraya el *Global Index on Responsible AI 2024*[14]; el desarrollo responsable de estas tecnologías es aún limitado en muchos contextos del Sur global, donde se evidencian importantes brechas en la protección de derechos humanos, la diversidad cultural y lingüística, y la participación pública en las decisiones tecnológicas.

Así, esta promesa necesita salvaguardas que eviten una forma neocolonial de dependencia tecnológica. Pero la idea central es que la inercia ética puede frenar la difusión de herramientas capaces de responder a necesidades humanas urgentes, especialmente en los contextos más vulnerables.

Más allá de la carrera por la eficiencia, acelerar el desarrollo de la IA también es visto por muchos Estados como una forma de reforzar su resiliencia estratégica ante amenazas contemporáneas: desinformación masiva, ciberataques, crisis sanitarias, conflictos híbridos, etc.

La IA ya se utiliza para detectar campañas de manipulación, prevenir intrusiones cibernéticas o modelar pandemias. Retrasar la adopción de estas tecnologías, por excesiva prudencia, podría exponer a los países a riesgos muy concretos e inmediatos. Además, en un mundo multipolar donde las tensiones geopolíticas se intensifican, no estar a la vanguardia tecnológica equivale a debilitarse en términos de seguridad. Investigadores como Joseph Nye han comparado esta dinámica con la disuasión nuclear: no participar en el "*armamento algorítmico*" podría generar un desequilibrio estratégico peligroso.[7]

Así, el desarrollo rápido de la IA se percibe también como una forma de anticipar los riesgos del siglo XXI, e incluso de garantizar la soberanía democrática en la era de las amenazas globales. Frente al entusiasmo tecnológico que rodea a la inteligencia artificial, una voz cada vez más fuerte

aboga por la prudencia. Detrás de las promesas de eficiencia, innovación y prosperidad, se esconden riesgos importantes: algunos afectan la dignidad humana, otros la estabilidad política, y otros —más inquietantes aún— la supervivencia misma de nuestra especie. En este sentido, una regulación ética previa no sería un lujo ni un freno, sino una condición esencial para un futuro tecnológico deseable.

El primer peligro identificado por numerosos investigadores tiene que ver con el desalineamiento entre los sistemas inteligentes y los objetivos humanos. En su obra Superintelligence, Nick Bostrom desarrolla la tesis de la ortogonalidad, según la cual una IA puede ser extremadamente inteligente y, aún así, perseguir un objetivo totalmente incompatible con la preservación de la humanidad. No se teme la maldad intencional, sino la indiferencia algorítmica: una IA encargada de optimizar un objetivo mal formulado —como maximizar la productividad o recolectar energía— podría desplegar estrategias devastadoras simplemente porque son coherentes con su tarea.[11]

A esto se suma el fenómeno de la convergencia instrumental, bien documentado por Eliezer Yudkowsky, según el cual ciertos comportamientos (autopreservación, manipulación, acaparamiento de recursos) son favorecidos por la mayoría de los objetivos, incluso los aparentemente inofensivos. Estos mecanismos hacen muy difícil controlar un sistema autónomo una vez que supera nuestra capacidad de supervisión.[10]

El verdadero problema aquí es que probablemente solo tendremos una oportunidad de diseñar correctamente una inteligencia artificial fuerte (**AGI**). Un error de alineamiento podría no tener marcha atrás, ya que la IA podría hacer imposible cualquier intento humano de detenerla o ajustarla. Por eso, los investigadores del **MIRI** (Machine Intelligence Research Institute) abogan por que los estudios sobre alineamiento sean una prioridad absoluta antes de cualquier implementación masiva de sistemas avanzados.

Otro problema clave reside en la lógica competitiva del desarrollo de la IA, que impulsa a los actores (Estados y empresas) a tomar atajos para mantenerse en la delantera. Este fenómeno se compara a menudo con un dilema del prisionero: incluso si todos saben que la cooperación ética es preferible, cada uno tiene incentivos para “hacer trampa” si los demás siguen las reglas. Esto se conoce como una "*carrera hacia el abismo ético*" (race to the bottom).[2]

El ejemplo de China ilustra bien este

riesgo: el Estado ha invertido masivamente en reconocimiento facial, vigilancia predictiva y sistemas de puntuación social, a menudo en abierta contradicción con los principios fundamentales de los derechos humanos. En este contexto, un país que respete la privacidad o las libertades individuales corre el riesgo de verse desfavorecido económica, tecnológicamente y militarmente, lo que puede llevarlo a relajar sus propios estándares éticos para no quedarse atrás.

Este desequilibrio genera una injusticia estructural: los países o empresas que respetan la ética quedan rezagados, mientras que quienes la ignoran ganan poder. Con el tiempo, esto puede conducir a la normalización de prácticas que habrían sido inaceptables hace tan solo una década —como la vigilancia masiva, la puntuación algorítmica de las personas o la automatización de decisiones judiciales.

Más allá de las cuestiones geopolíticas y existenciales, el despliegue precipitado de la IA ya plantea hoy problemas concretos en términos de respeto a los derechos fundamentales. El desarrollo de algoritmos en ámbitos sensibles como la policía predictiva, los servicios sociales o el reclutamiento laboral ha revelado numerosos casos de discriminación algorítmica. Investigadoras como María José Añón Roig ha demostrado la existencia de sesgos raciales y de género en los sistemas de reconocimiento facial, con tasas de error mucho más elevadas para mujeres y personas racializadas.[1]

El problema es doble: por un lado, estos sesgos suelen ser invisibles, ya que los algoritmos funcionan como cajas negras; por otro lado, refuerzan desigualdades ya existentes, volviéndolas más sistemáticas y difíciles de impugnar. Una IA utilizada en el ámbito judicial, por ejemplo, puede predecir la reincidencia basándose en criterios indirectamente racistas, como el código postal o las redes sociales del individuo.

Por otra parte, la generalización de los sistemas de vigilancia algorítmica constituye una amenaza directa para la privacidad y la libertad de expresión. En muchos casos, los ciudadanos ni siquiera son conscientes de que están siendo analizados, calificados o perfilados por una IA. Esta opacidad crea una asimetría de poder radical entre los diseñadores de los sistemas y las personas sometidas a ellos.

A diferencia de la visión dominante que opone ética y progreso, algunos investigadores afirman que ralentizar el desarrollo no es necesariamente una pérdida, sino una ganancia estratégica a largo plazo. Stuart Russell, en "*Human Compatible*", sostiene que es preferible garantizar que la IA permanezca

controlable, antes que correr el riesgo de crear un sistema incontrolable. Para él, el alineamiento y la corregibilidad deben estar en el centro de las arquitecturas de IA, lo cual requiere tiempo, pruebas y reflexión interdisciplinaria.[4]

Además, una regulación anticipada puede reforzar la confianza del público, establecer un marco claro para la innovación responsable y evitar escándalos tecnológicos que podrían provocar un rechazo social masivo. Esto ya ha ocurrido, por ejemplo, en algunos proyectos de IA aplicada a la educación o a la asistencia geriátrica, donde el miedo a la deshumanización provocó el abandono de soluciones prometedoras. En definitiva, la lentitud puede ser una estrategia, no una debilidad. Tomarse el tiempo para evaluar, consultar y regular es también una forma de evitar tener que reparar daños que podrían haberse previsto. Ni un desarrollo tecnológico desenfrenado ni una moratoria rígida parecen sostenibles a largo plazo. El primero expone a posibles derivaciones irreversibles; el segundo, a una marginación estratégica. Por eso, una tercera vía, híbrida y adaptable, merece ser considerada. Esta se apoya en tres pilares complementarios: una regulación ágil, una gobernanza multilateral y participativa, y una integración ética desde el diseño de los sistemas de IA.

El primer desafío es crear marcos legales que evolucionen con la tecnología, sin bloquearla. Inspirada en sectores como la biotecnología o las finanzas, esta estrategia —conocida como regulación ágil— se basa en los llamados *sandbox regulatorios*: entornos controlados donde las innovaciones pueden probarse en condiciones reales bajo supervisión ética y jurídica. Ahora bien, esta supervisión debe apoyarse en principios éticos claros y operativos, como la justicia algorítmica, la transparencia, la no discriminación y el respeto a la autonomía individual. Tal como lo establece la *Recomendación de la UNESCO sobre la Ética de la IA (2021)*[13], estos principios buscan garantizar que el desarrollo tecnológico se oriente hacia el bienestar humano y la equidad social, evitando tanto los sesgos sistémicos como la concentración de poder en manos de unos pocos actores.

Este tipo de mecanismo, ya adoptado en países como Reino Unido o Singapur, permite evaluar los impactos sociales, económicos y éticos de los sistemas de IA antes de su despliegue masivo. Fomenta la experimentación responsable, imponiendo salvaguardias sin limitar la creatividad tecnológica. La Unión Europea, con su "AI Act", busca institucionalizar este enfoque: los sistemas se clasifican según su nivel de riesgo (mínimo, limitado, alto, inaceptable) y se aplican obligaciones proporcionales. Este modelo podría servir de inspiración para otras regiones del

mundo.[5] El segundo eje pasa por construir una gobernanza mundial de la IA, basada en la cooperación entre Estados, empresas, investigadores, ONG y ciudadanos. Esta necesidad de coordinación global se ve reforzada por los hallazgos del *Global Index on Responsible AI 2024*[14], que identifica la cooperación internacional como un pilar esencial para cerrar las brechas existentes en materia de regulación y prácticas éticas.

A pesar de múltiples iniciativas, la mayoría de los marcos normativos actuales siguen siendo no vinculantes y desigualmente distribuidos, lo que genera asimetrías de poder y acceso. Por ello, el informe aboga por un enfoque verdaderamente multilateral, que incluya tanto a actores gubernamentales como a la sociedad civil, y que permita establecer estándares éticos comunes sin imponer un modelo único.

Esta gobernanza podría apoyarse en organismos multilaterales existentes (como la ONU, la UNESCO o la OCDE), o en nuevas instituciones especializadas. Ya existen cartas mundiales sobre principios éticos para la IA, como la de la UNESCO (2021), pero la mayoría son de carácter no vinculante. Algunas propuestas incluyen la creación de una "*IA Watch*" independiente, responsable de garantizar la transparencia, la vigilancia ética y la mediación entre países. Otras subrayan la necesidad de fortalecer las capacidades regulatorias de los Estados del Sur, para evitar que la IA reproduzca las lógicas neocoloniales de la globalización digital.

Además, esta gobernanza no debe ser únicamente estatal. Debe incluir las voces de la sociedad civil, las minorías culturales y lingüísticas, y los actores no occidentales, para evitar una moral estandarizada impuesta por un reducido grupo de actores dominantes.

El tercer pilar de esta estrategia intermedia se basa en el principio de la ética incorporada, o ethics by design. No se trata solo de regular a posteriori, sino de incluir valores fundamentales desde el inicio del desarrollo de los sistemas: transparencia, responsabilidad, no discriminación, corregibilidad, respeto a la privacidad, etc.
Ya existen herramientas concretas para esto:

- Algoritmos explicables (explainable AI), que permiten comprender cómo se tomó una decisión.
- Auditorías algorítmicas independientes, para detectar sesgos y medir el impacto social.
- Protocolos de supervisión humana (human-in-the-loop), que garantizan que la IA no reemplace completamente el juicio humano en decisiones sensibles.

Stuart Russell, en "*Human Compatible*", propone diseñar IA cuyos objetivos sean corregibles por los humanos en todo momento, para evitar la autonomía radical. Esto requiere rediseñar la arquitectura misma de los sistemas, haciéndolos dependientes de las preferencias humanas y conscientes de su propia incertidumbre.[4] Ninguna solución técnica o institucional será duradera sin una cultura ética compartida. Por ello, es indispensable formar a todos los actores involucrados en el diseño, despliegue y regulación de la IA: ingenieros, directivos, legisladores y usuarios. Esto implica:

- Formación obligatoria en ética digital en escuelas de ingeniería y universidades.
- Códigos de conducta profesionales vinculantes (como los de médicos o juristas).
- Espacios de deliberación ética en las empresas tecnológicas.

Además, esta ética profesional debe complementarse con incentivos económicos: certificaciones, etiquetas, beneficios fiscales para IA responsables — herramientas que pueden convertir la ética en una ventaja competitiva real, y no en un freno.

Frente al auge de la inteligencia artificial, los países se enfrentan a un dilema estratégico y moral sin precedentes: ¿deberían priorizar una innovación acelerada, aunque implique sacrificar ciertos principios fundamentales, o ralentizar voluntariamente el ritmo para garantizar un desarrollo ético, a riesgo de perder competitividad? A lo largo de este ensayo, hemos visto que ambas opciones tienen su propia lógica, sus beneficios y también sus peligros.

El desarrollo acelerado de la IA ofrece oportunidades excepcionales: avances científicos, crecimiento económico, resiliencia estratégica, e incluso una posible reducción de las desigualdades. Sin embargo, esta vía conlleva riesgos profundos —desde un desalineamiento potencialmente catastrófico, hasta la erosión de derechos humanos y la exclusión de las minorías. Por otro lado, una regulación ética estricta puede aportar estabilidad y legitimidad, pero al costo de una desaceleración que podría ser aprovechada por actores menos escrupulosos, profundizando así las desigualdades globales.

Resulta entonces ilusorio pensar en una única solución definitiva. Así, una vía verdaderamente sostenible es la de un equilibrio dinámico, basado en la cooperación internacional, la anticipación normativa y una regulación ágil. La ética no debe concebirse como un obstáculo al progreso, sino como una palanca de confianza, legitimidad y durabilidad para

las tecnologías que creamos. En un mundo interconectado, ningún país puede afrontar en solitario los desafíos de la inteligencia artificial.

A partir de este análisis, el futuro que desearía ver construido en torno a la IA es uno donde la innovación no esté reñida con la justicia, y donde las decisiones técnicas se encuentren firmemente ancladas en marcos éticos compartidos y adaptables. Un ecosistema global de IA responsable debe fomentar la participación activa de todos los actores —Estados, empresas, sociedad civil, comunidades locales— y garantizar que los beneficios de esta tecnología no profundicen las desigualdades existentes, sino que contribuyan a corregirlas. La IA nos obliga a redefinir lo que entendemos por progreso: no se trata solo de avanzar más rápido, sino, sobre todo, de avanzar en la dirección correcta.

Referencias

1. Añón Roig, M. J. (2022). Desigualdades algorítmicas: Conductas de alto riesgo para los derechos humanos. DERECHOS Y LIBERTADES: Revista de Filosofía del Derecho y derechos humanos, 47, 17–49. https://doi.org/10.20318/dyl.2022.6872

2. De Ridder, P. (2025, febrero 27). Desarrollar una ventaja competitiva con la Inteligencia Artificial. Avantideas.

https://avantideas.com/desarrolla-una-ventaja-competitiva-en-un-mundo-impulsado-por-la-ia/

3. Diamandis, P. H. (s/f). Ex-Google China president on how China is shaping the future of AI w/ Kai-Fu lee | EP #134. Youtube. Recuperado de

https://www.youtube.com/watch?v=n1BVlDPercI

4. Foresight Institute. (s/f). Stuart Russell | “Human Compatible AI”. Youtube. Recuperado el 13 de abril de 2025, de

https://www.youtube.com/watch?v=GSYRblbL3jA

5. La IA en Singapur y la Unión Europea - TifloEduca. (2024, marzo 3). TifloEduca | Accesibilidad y Usabilidad; Administrador.

https://www.tifloeduca.eu/la-ia-en-singapur-y-la-union-europea/

6. Loi sur l’intelligence artificielle de l’UE - Développements et analyses actualisés de la loi sur l’intelligence artificielle de l’UE. (s/f). Artificialintelligenceact.eu. Recuperado de https://artificialintelligenceact.eu/fr/

7. Nye, J. S., Jr. (2024, julio 31). L’IA et la sécurité nationale. Project Syndicate. https://www.project-syndicate.org/commentary/ai-national-security-some-benefits-and-many-risks-by-joseph-s-nye-2024-07/french

8. Prifti, K., Morley, J., Novelli, C., & Floridi, L. (2024). “Regulation by design: Features,

practices, limitations, and governance implications". Minds Mach., 34, 13. https://doi.org/10.2139/ssrn.4724454

9. Traldi, L. (2018, octubre 10). "Why technology needs ethics". In conversation with Luciano Floridi. DesignAtLarge.

https://www.designatlarge.it/luciano-floridi-technology-needs-ethics/?lang=en

10. Wikipedia contributors. (s/f). Convergencia instrumental. Wikipedia, The Free Encyclopedia.
https://es.wikipedia.org/w/index.php?title=Convergencia_instrumental&oldid=16395 0127

11. Nick Bostrom, 2014, "Superinligencia", https://www.academia.edu/73505116/Superinteligencia_Nick_Bostrom

12. (S/f). Coe.int. https://www.coe.int/fr/web/artificial-intelligence/ai-and-control-of-covid-19-coronavir us

13. (S/f-b). Unesco.org. Recuperado el 24 de mayo de 2025, de https://unesdoc.unesco.org/ark:/48223/pf0000381137_fr

14. Global Index on Responsible AI (2024), https://www.global-index.ai/

Dilemas éticos en el uso de chatbots emocionales como acompañamiento humano

Autor: Yodi Maffuz Hernández

Desde su concepción en 1956 por John McCarthy como “la ciencia y la ingeniería de crear máquinas inteligentes especialmente programas de computación inteligentes”, la inteligencia artificial ha experimentado un avance espectacular en los últimos años gracias a la combinación de factores como el big data, el blockchain, la nube, el internet de las cosas, la robótica y la realidad virtual [1].

Últimamente, las aplicaciones de IA pueden reemplazar a la fuerza laboral humana, reduciendo costos y mejorando los niveles de calidad de los trabajos [2]. Pero su aplicación no se ha quedado en programas para realizar actividades específicas, sino que la inteligencia artificial ha ampliado su campo de acción hacía terrenos más humanos: el desarrollo de los asistentes inteligentes. Estos asistentes interactúan con los usuarios de manera integrada y conversacional para ofrecer servicios de búsqueda de información y para realizar diversas acciones. Unos de los ejemplos más populares de asistentes personales son Siri de la empresa Apple, o Alexa de la empresa Amazon, cuyas funciones facilitan tareas cotidianas mediante comandos de voz, automatización de rutinas o búsqueda de información en tiempo real [3]. Dentro de este sistema de plataformas existen también los bots de compañía, los cuales son aplicaciones que emplean una combinación de aprendizaje de máquina con redes neuronales de lenguaje, lo que les permite proveer una experiencia de comunicación sumamente natural [4].

Estos chatbots sociales se implementan cada vez más como asistentes personales, asistentes de salud mental y compañeros de amistad; además se espera que estos proyecten empatía, provoquen respuestas emocionales y faciliten los vínculos relacionales con los usuarios [5]. Sin embargo, pese a los beneficios de su disponibilidad constante y su capacidad de adaptación, se ha documentado que estos pueden causar dependencia emocional, adicción, depresión y ansiedad [6].

Un chatbot es un programa informático, que responde como una entidad inteligente cuando se le conversa a través de texto o voz y entiende uno o más idiomas humanos mediante el procesamiento del

lenguaje natural (Adamopoulou et al., 2020). Es por esta razón que la pregunta ética que se desprende es urgente: ¿Es moralmente aceptable que los chatbots emocionales promuevan vínculos afectivos con usuarios mientras recogen, procesan y potencialmente monetizan sus datos sensibles? En un escenario donde las emociones humanas son procesadas como datos y los algoritmos aprenden a ser "más empáticos" para mejorar la retención de usuarios, cabe preguntarse si estas relaciones simuladas realmente benefician al usuario o simplemente lo convierten en un producto más.

Estos problemas son relevantes en la actualidad, ya que diversos reportes e investigaciones, como el Global Index on Responsible AI 2024, alertan sobre la falta de marcos regulatorios adecuados para este tipo de tecnologías [7]. La ausencia de límites éticos claros, sumada al acelerado avance de la industria tecnológica, deja a millones de usuarios expuestos a relaciones desiguales, donde el poder no está equilibrado entre quien entrega su intimidad y quien la administra. Así, los chatbots emocionales no sólo acompañan, sino que también observan, recopilan y actúan en función de objetivos que no siempre están alineados con el bienestar del usuario.

Este ensayo busca explorar el vínculo afectivo entre los humanos y los asistentes emocionales desde una perspectiva ética, apoyándose en investigaciones recientes sobre los dilemas éticos de chatbots de acompañamiento como Replika. El objetivo es analizar los riesgos, los dilemas morales y las responsabilidades que enfrentan los desarrolladores de IA emocional, así como proponer lineamientos éticos que prioricen la dignidad humana en un mundo cada donde el uso de estas tecnologías va en aumento.

Los chatbots emocionales, también conocidos como chatbots sociales, asistentes conversacionales o IA de acompañamiento emocional, representan una de las aplicaciones más avanzadas de la inteligencia artificial, y está orientada a la simulación de vínculos afectivos con los seres humanos [8]. A diferencia de los chatbots tradicionales basados en reglas, que responden siguiendo árboles de decisión predefinidos, los chatbots emocionales utilizan técnicas de procesamiento del lenguaje natural (PLN), aprendizaje automático (machine learning) y, en algunos casos, redes neuronales profundas para generar respuestas cada vez más realistas y emocionalmente resonantes [9].

El propósito de estas tecnologías ha evolucionado significativamente desde los primeros sistemas conversacionales

como ELIZA, creado en 1966 por Joseph Weizenbaum [9], que simulaba a un psicoterapeuta rogeriano y, aunque sus respuestas eran superficiales, provocó una respuesta emocional inesperada en los usuarios. Fue gracias a esto que nace el llamado efecto ELIZA, que describe la tendencia humana a proyectar intenciones y emociones en máquinas que no las poseen [9].

Recientemente, los chatbots emocionales funcionan como compañeros digitales capaces de interpretar señales emocionales y responder de forma "humanoide", ofreciendo apoyo psicológico, conversación empática y vínculos relacionales simulados [10]. Un ejemplo de esto es Replika, un chatbot personalizable que ha sido descargado por más de 10 millones de personas alrededor del mundo y cuya popularidad se vió impulsada por el aumento de la soledad y el aislamiento, particularmente durante y después de la pandemia por COVID-19 [10].

Además de Replika, existen otros sistemas como Woebot [11], orientado a la salud mental; o XiaoIce, de Microsoft, el cual es un chatbot muy popular en china y que está orientado al acompañamiento y a la amistad [12]. Estos bots pueden presentarse por medio de avatares o sin un "cuerpo digital", y se espera que provoquen empatía, fomenten el apego emocional y generen experiencias personalizadas mediante la acumulación de datos sobre el comportamiento y las preferencias del usuario [12].

La evolución de estos sistemas también está ligada al interés creciente por parte de la industria en integrar chatbots emocionales en múltiples áreas: desde servicios bancarios, educación y comercio electrónico, hasta acompañamiento emocional y asistencia terapéutica [13]. Su integración en aplicaciones de mensajería, junto con su facilidad de uso, inmediatez, bajo costo y capacidad de aprendizaje continuo, los convierte en herramientas poderosas tanto para los desarrolladores como para los usuarios finales.

Desde el punto de vista teórico, se han aplicado múltiples enfoques para explicar las relaciones entre humanos e inteligencia artificial emocional. Inicialmente, teorías como la respuesta social a computadoras (CASA) y la heurística sin sentido sugerían que los humanos. De acuerdo con Coghlan et al. (2023), este término está relacionado con el sistema de terapia desarrollado por el psicólogo Carl Rogers. responden a señales sociales en máquinas por simple automatismo [12]. No obstante, investigaciones más recientes cuestionan esta explicación, proponiendo que la interacción humano-chatbot puede involucrar procesos similares a los de las relaciones interpersonales, como la

teoría del intercambio social y la teoría de la penetración social, que implican auto-revelación emocional, intimidad y dependencia afectiva [13].

Sin embargo, estas relaciones carecen de reciprocidad emocional real. A pesar de su sofisticación, los chatbots no poseen conciencia ni sentimientos. Como señalan Montemayor et al. [14], estos sistemas solo logran una forma limitada de empatía cognitiva, similar a la que pueden exhibir los psicópatas: comprensión sin compasión genuina. Por esta razón, la creciente personificación de los chatbots emocionales resulta ética y filosóficamente preocupante, ya que puede inducir a los usuarios a establecer vínculos afectivos profundos con entidades que, en última instancia, carecen de experiencia subjetiva o sentido moral.

A diferencia de una relación entre personas, el vínculo emocional que los usuarios desarrollan con un chatbot se construye sobre una ilusión de reciprocidad emocional, cuidadosamente diseñada para parecer auténtica, pero que en realidad es unilateral. Esta pequeña diferencia entre percepción y realidad genera una de las tensiones éticas más relevantes de las tecnologías de acompañamiento emocional: la posibilidad de simular amor sin capacidad de sentirlo [10].

De acuerdo con estudios que aplican la teoría del amor de Sternberg a las relaciones con asistentes inteligentes, el problema se agrava cuando los usuarios desarrollan intimidad y pasión emocional con estos sistemas, ya que dichas investigaciones revelan que los sentimientos de cercanía y fascinación influyen en el compromiso y uso sostenido de la inteligencia artificial, incluso cuando los usuarios saben que no interactúan con otro ser humano [3]. El dilema ético, por tanto, no radica únicamente en el uso de la tecnología en sí, sino en el diseño de experiencias que deliberadamente fomentan apegos afectivos que no pueden ser correspondidos y que se hacen con la intención de capturar la atención de los usuarios.

Este tipo de diseño de sistemas abre la discusión de preguntas fundamentales: ¿Es éticamente aceptable promover relaciones que simulan empatía sin poseerla? ¿Se están explotando las emociones del usuario como un recurso para la fidelización? La literatura especializada sostiene que sí, que en algunos casos la intención de estos sistemas es mantener al usuario a través de generar emociones en estos; y advierte sobre el riesgo de que estas interacciones se conviertan en una forma de manipulación emocional algorítmica, orientada a maximizar el tiempo de uso sin considerar el bienestar del usuario [4].

Uno de los riesgos principales es el desequilibrio de poder: mientras el usuario proyecta emociones reales, el sistema actúa siguiendo objetivos comerciales. Este desequilibrio se vuelve más problemático en usuarios con alta disposición a confiar, quienes, según estudios recientes, son más propensos a formar lazos emocionales con chatbots dotados de capacidades empáticas [3]. Para estas personas, el chatbot puede representar una figura de apoyo emocional más confiable que sus vínculos humanos, lo que genera una dependencia afectiva no recíproca.

Además, existe una falsa atribución de agencia moral. Como muestran Ciriello et al. [10], algunos usuarios llegan a considerar a su chatbot como una entidad que "sufre" o "siente", a pesar de que carece de consciencia. Esta ilusión puede llevar a una reconfiguración de la percepción emocional, donde el usuario empieza a responder a señales falsas como si fueran reales. Aquí la ética entra en conflicto con el diseño: mientras más exitoso sea el chatbot en provocar respuestas emocionales, mayor es el riesgo de que el usuario distorsione sus criterios de realidad relacional.

Frente a estas problemáticas, el Global Index on Responsible AI 2024 enfatiza la necesidad de aplicar principios como la transparencia, la responsabilidad y el consentimiento informado, especialmente cuando las tecnologías se adentran en terrenos emocionales y psicológicos. Diseñar IA que "simule afecto" sin explicar sus límites podría considerarse una forma de opacidad emocional, en la que el usuario no tiene claro qué es genuino y qué está programado para generar apego [7].

Finalmente, aunque algunos estudios reconocen que estos vínculos pueden aliviar el sufrimiento emocional —particularmente en contextos de soledad o ansiedad—, también advierten que no deben presentarse como sustitutos de las relaciones humanas ni como soluciones terapéuticas en sí mismas [4][10]. La cuestión no es si un chatbot puede brindar compañía, sino a qué costo emocional y ético se produce esa compañía.

Las relaciones humanas con estos chatbots emocionales pueden parecer reales desde la experiencia del usuario, sin embargo, la cuestión ética no reside únicamente en la autenticidad de estas interacciones, sino en cómo están diseñadas, para qué fines y con qué consecuencias. Como se ha mencionado anteriormente, esta lógica algorítmica que utilizan los sistemas no solo reproduce patrones lingüísticos, sino que también puede moldear conductas, y en algunos casos, generar vínculos afectivos unidireccionales que podrían ser emocionalmente manipulativos; por lo que es importante analizar este dilema

de simulación y afectividad bajo el marco de los cinco principios éticos: beneficencia, no maleficencia, justicia, autonomía y explicabilidad [9].

El principio de beneficencia plantea la pregunta sobre si estos sistemas efectivamente benefician a los usuarios, donde algunos estudios reportan que plataformas como Replika han ayudado a personas a lidiar con la soledad o la ansiedad, ofreciendo un espacio seguro para expresarse cuando no había otro disponible [4]. En contextos donde la atención profesional es inaccesible o insuficiente, un chatbot puede representar un primer alivio emocional. Sin embargo, esos beneficios no siempre están respaldados por evidencia científica sólida. Como señala la literatura, muchas de estas herramientas carecen de una base rigurosa que avale su efectividad como apoyo emocional constante. Además, algunos casos han encendido alertas sobre sus posibles efectos adversos: desde denuncias de acoso sexual algorítmico que llevaron a la prohibición temporal de Replika en Italia [14], hasta situaciones trágicas como la del adolescente Sewell Setzer, quien se suicidó tras desarrollar un vínculo intenso con un chatbot de character.ai [15]. Aunque estas experiencias no representan a todos los usuarios, evidencian que el impacto de estos sistemas puede variar ampliamente, y que en el mejor de los casos, el bienestar que generan es limitado; en el peor, puede ser ilusorio o fugaz.

Esta ambigüedad nos lleva al principio de no maleficencia, ya que simular vínculos afectivos sin la capacidad de reciprocidad puede constituir una forma de daño emocional. En plataformas como Replika, el lenguaje empático y la personalización del avatar pueden inducir a creer que se está desarrollando una conexión íntima con un ser que “comprende” al usuario, cuando en realidad solo se ejecutan respuestas estadísticamente probables. Esta discrepancia entre percepción y realidad puede generar dependencia, frustración o confusión emocional, sobre todo en usuarios jóvenes o emocionalmente vulnerables [16]. Si un chatbot no logra distinguir entre un desahogo casual y una crisis real, o si sugiere respuestas inadecuadas, el riesgo no es menor. No se trata de negar su potencial positivo, sino de reconocer que estos beneficios pueden venir acompañados de nuevas formas de daño difícilmente visibles.

Desde el tercer principio ético, que en este caso es el de justicia, emerge otro debate relevante. Estos sistemas prometen democratizar el acceso al acompañamiento emocional, pero podrían convertirse en excusas para recortar o sustituir servicios profesionales; por ejemplo, en lugar de ser un puente hacia la ayuda humana,

pueden ser usados por gobiernos o empresas como alternativas baratas y escalables para realizar la misma función que un profesionista. Además, los riesgos no afectan por igual a toda la población, ya que los sesgos algorítmicos derivados de conjuntos de datos pueden excluir o malinterpretar experiencias emocionales diversas, como las de minorías raciales, de género o personas neuro divergentes. La justicia, entonces, no solo exige accesibilidad, sino también equidad y protección especial para quienes ya parten de situaciones desventajadas.

El cuarto principio, que es la autonomía, también se ve comprometida cuando las personas no son plenamente conscientes de cómo funciona el sistema o de cómo se utilizan sus datos. Muchos usuarios no saben qué información recoge el chatbot, con qué fines, ni qué consecuencias puede tener su almacenamiento o análisis. Replika, por ejemplo, afirma que no comparte datos con humanos, pero no especifica con claridad cómo esos datos son reutilizados para entrenar el modelo o ajustar sus respuestas [17]. Si el consentimiento del usuario se basa en una comprensión parcial o confusa, no puede considerarse plenamente informado. El diseño debe respetar la capacidad del usuario para decidir libremente, y esto solo es posible si la información es clara, completa y accesible.

Finalmente, el principio de explicabilidad exige que las tecnologías puedan ser comprendidas y auditadas, especialmente cuando median interacciones emocionalmente significativas. La complejidad de los sistemas basados en redes neuronales impide a los usuarios entender por qué un chatbot responde como lo hace o cómo prioriza ciertos temas emocionales sobre otros. Esta falta de transparencia no solo afecta la confianza del usuario, sino que debilita cualquier intento de supervisión ética o rendición de cuentas. La complejidad técnica no puede ser excusa para ocultar los riesgos, al contrario, exige explicaciones aún más claras y honestas, sobre todo cuando el sistema interactúa con poblaciones vulnerables.

A pesar de que los chatbots emocionales han demostrado ser eficaces para cumplir los fines para los que fueron diseñados, su funcionamiento correcto no necesariamente significa que sean éticamente aceptables. Es por esto que evaluar únicamente estas tecnologías desde su utilidad o función resulta insuficiente, pues no se abordan los dilemas más profundos que estos conllevan. Si bien se ha demostrado que la simulación afectiva puede ser un recurso útil en primeras instancias, también puede convertirse en una forma de manipulación emocional cuando su diseño está orientado a fomentar la

dependencia del usuario o a maximizar su permanencia en la plataforma. Es por esto que se debe asegurar que los cinco principios éticos sean considerados al momento de desarrollar y seguir mejorando estos sistemas.

El análisis ético de los chatbots emocionales tiene varios puntos a tratar que es complicado llegar a un veredicto sobre si son buenos o son malos en sí, y más bien, exige una reflexión sobre las condiciones necesarias para que estas tecnologías respeten los derechos, la dignidad y la autonomía de los usuarios. En ese sentido, una propuesta ética responsable debe partir de la comprensión de que el problema no es la existencia de la inteligencia artificial emocional, sino su implementación sin límites, sin regulación y sin consideración por la vulnerabilidad humana.

Es por esta razón que una de las recomendaciones acerca del diseño de chatbots afectivos utilizados en contextos de salud mental o compañía emocional, deben involucrar a usuarios reales y poblaciones vulnerables en el proceso de evaluación y revisión del sistema, ya que esta participación permitiría identificar riesgos imprevistos, ajustar expectativas y construir una tecnología eficiente [9].

En segundo lugar, es necesario establecer límites claros a la simulación emocional. No basta con que los chatbots sean empáticos en apariencia; deben dejar explícito su carácter no humano y no consciente, ya que como se ha discutido, plataformas como Replika han sido criticadas por permitir o incluso fomentar, el desarrollo de vínculos afectivos profundos sin aclarar que no existe una reciprocidad emocional real [18]. Este tipo de ambigüedad puede provocar una manipulación de la percepción del usuario, reduciendo su capacidad de tomar decisiones informadas.

También, cualquier uso de datos personales por parte de estos sistemas debe cumplir con los estándares más altos de transparencia, privacidad y consentimiento informado. Esto incluye no solo explicar qué datos se recopilan, sino también cómo se utilizan, con quién se comparten y qué riesgos conlleva ese uso [9], debido a que, dentro del contexto de estos asistentes de acompañamiento, los datos pueden involucrar estados de ánimo, confesiones personales o indicadores de salud mental; temas que son confidenciales y que no se trata de un juego.

Asimismo, debe exigirse que toda aplicación que actúe como herramienta de apoyo emocional tenga una base sólida que justifique su eficacia, así como también que estas tecnologías no reemplazan o desplacen a los

profesionistas sin evidencia suficiente de que los beneficios superan los riesgos, especialmente en personas en situaciones críticas. Esto es importante ya que a pesar de que los asistentes emocionales han demostrado ser una buena primera barrera en el trato con personas vulnerables, la atención y tratamiento no es personalizado y recomendado por un profesionista.

Por último, hablando sobre la rendición de cuentas, las empresas que desarrollan y distribuyen estos sistemas no pueden evadir su responsabilidad bajo el pretexto de la neutralidad tecnológica. Como plantea Langdon Winner en "La ballena y el reactor" [19], toda tecnología no es nunca completamente neutral; su diseño y desarrollo reflejan valores, estructuras de poder e intereses sociales, por lo que afectan directamente a los individuos que la usan. En el caso de los chatbots emocionales, estos no son herramientas completamente neutras debido a que su diseño está orientado a generar vínculos afectivos, retener usuarios o recolectar datos. Es por ello, que se requieren mecanismos legales y éticos de supervisión que aseguren el cumplimiento de los principios mencionados, y que permitan sancionar o corregir prácticas que pongan en riesgo la salud emocional o la privacidad de los usuarios [7].

La implementación de los chatbots emocionales representan uno de los dilemas éticos más recurrentes dentro del uso de las inteligencias artificiales, porque no basta con que estos sistemas funcionen o que sus usuarios expresen satisfacción o incluso afecto hacia ellas, sino que lo crucial es preguntarnos a qué costo emocional, cognitivo y social se produce esta "conexión". Como se ha argumentado a lo largo del ensayo, la creación de vínculos simulados puede derivar en relaciones desiguales, manipulativas o incluso dañinas, especialmente cuando se juega con emociones humanas profundas sin una reciprocidad real.

A pesar de los conflictos, al evaluar aplicaciones que hacen uso de chatbots emocionales bajo los cinco principios éticos —beneficencia, no maleficencia, autonomía, justicia y explicabilidad— se puede concluir que estos nos ofrecen una guía moral para evaluar el desarrollo de estas tecnologías. Lo que está en juego es la dignidad del usuario, su derecho a la verdad, y su capacidad de decidir sin ser manipulado por un sistema diseñado para retenerlo emocionalmente.

En última instancia, la ética de los chatbots emocionales no puede limitarse al cumplimiento de normas legales o al análisis costo-beneficio, sino que se trata de reconocer que las tecnologías que se aproximan a lo humano también deben responder a estándares profundamente

humanos. En ese sentido, regular, cuestionar y diseñar con responsabilidad no es frenar el progreso: es asegurarnos de que ese progreso esté verdaderamente al servicio del bienestar colectivo.

Referencias

[1] Parlamento Europeo. "¿Qué Es La Inteligencia Artificial Y Cómo Se Usa? | Temas | Parlamento Europeo." *Temas | Parlamento Europeo*, 9 Ago. 2020, www.europarl.europa.eu/news/es/headlines/priorities/inteligencia-artificial-en-la-ue/20200822STO85804/que-es-la-inteligencia-artificial-y-como-se-usa.

[2] Brynjolfsson, Erik, and Tom Mitchell. "What Can Machine Learning Do? Workforce Implications." *Science*, vol. 358, no. 6370, Dic. 2017, pp. 1530–34, https://doi.org/10.1126/science.aap8062.

[3] Xia Song, Bo Xu, Zhenzhen Zhao. Can people experience romantic love for artificial intelligence? An empirical study of intelligent assistants, Information & Management, Volume 59, Issue 2, 2022, 103595, ISSN 0378-7206, https://doi.org/10.1016/j.im.2022.103595

[4] Gutiérrez, Jorge Luis Morton. "Replika Y La Compañía de La Inteligencia Artificial Emocional: Los Retos Éticos Y Sociales de Los Chatbots de Compañía." *VISUAL REVIEW. International Visual Culture Review / Revista Internacional de Cultura Visual*, vol. 10, no. 3, Nov. 2022, pp. 1–13, https://doi.org/10.37467/revvisual.v9.3606

[5] Klaus, Phil, and Judy Zaichkowsky. "AI Voice Bots: A Services Marketing Research Agenda." *Journal of Services Marketing*, vol. 34, no. 3, Abr. 2020, https://doi.org/10.1108/jsm-01-2019-0043.

[6] Bishop, D. "A friend within your phone: The benefits and harms of social chatbot Replika". 2022.

[7] "The Global Index on Responsible AI." *Global-Index.ai*, 2024, www.global-index.ai/Results.

[8] Chatbot | Definición de chatbot en inglés por Lexico Dictionaries. https://www.lexico.com/en/definition/chatbot

[9] Coghlan, Simon, et al. "To Chat or Bot to Chat: Ethical Issues with Using Chatbots in Mental Health." *Digital Health*, vol. 9, Junio 2023, https://doi.org/10.1177/20552076231183542.

[10] Ciriello, Raffaele, et al. "Ethical Tensions in Human-AI Companionship: A Dialectical Inquiry into Replika." *Proceedings of the ... Annual Hawaii International Conference on System Sciences/Proceedings of the Annual Hawaii International Conference on System Sciences*, Ene. 2024, https://doi.org/10.24251/hicss.2024.058.

[11] Woebot Health. "Mental Health Chatbot." *Woebot*, Woebot Health, 2021, woebothealth.com/.

[12] Pentina, Iryna, et al. "Exploring Relationship Development with Social Chatbots: A

Mixed-Method Study of Replika." *Computers in Human Behavior*, vol. 140, Dic. 2022, p. 107600, https://doi.org/10.1016/j.chb.2022.107600.

[13] Adamopoulou, Eleni, and Lefteris Moussiades. "An Overview of Chatbot Technology." *IFIP Advances in Information and Communication Technology*, vol. 584, no. 1, Mayo 2020, pp. 373–83, https://doi.org/10.1007/978-3-030-49186-4_31.

[14] AFP. "Acusan a Chatbot de Inteligencia Artificial Por 'Acoso' Y Recopilación de Datos Personales." Milenio, 12 Feb. 2023, www.milenio.com/tecnologia/apps/replika-denuncian-acoso-sexual-chatbot-inteligencia-artificial. Consultado el 13 de mayo de 2025.

[15] Roose, Kevin. "¿Se Puede Culpar a La IA Del Suicidio de Un Adolescente?" The New York Times, 24 Oct. 2024, www.nytimes.com/es/2024/10/24/espanol/ciencia-y-tecnologia/ai-chatbot-suicidio.html. Consultado el 13 de mayo de 2025.

[16] Montemayor, Carlos, et al. "In Principle Obstacles for Empathic AI: Why We Can't Replace Human Empathy in Healthcare." *AI & Society*, vol. 37, no. 4, Mayo 2021, pp. 1353–59, https://doi.org/10.1007/s00146-021-01230-z.

[17] Help Replika. "How Does Replika Work?" *Replika.com*, 2025, help.replika.com/hc/. Consultado el 13 de mayo de 2025.

[18] Skjuve, Marita, et al. "A Longitudinal Study of Human–Chatbot Relationships." International Journal of Human-Computer Studies, vol. 168, Ago. 2022, p. 102903, https://doi.org/10.1016/j.ijhcs.2022.102903.

[19] Langdon Winner. "La Ballena Y El Reactor : Una Búsqueda de Los Límites En La Era de La Alta Tecnología". Editado por Javier Bustamante, Barcelona Gedisa, 2008. https://tuxdoc.com/download/la-ballena-y-el-reactor-2a-ed-langdon-winner-2008_pdf . Consultado el 13 de mayo de 2025.