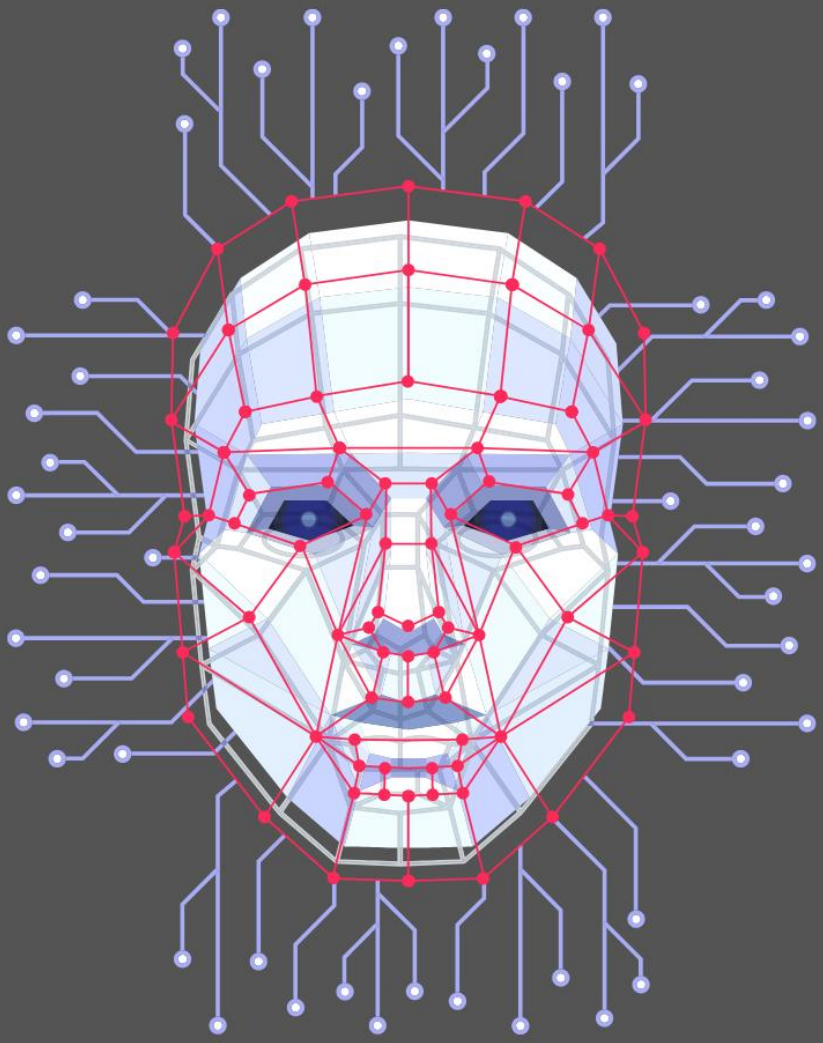
CONSEJO DE CIENCIA Y TECNOLOGÍA DEL ESTADO DE PUEBLA



PRINCIPIOS NORMATIVOS PARA UNA ÉTICA DE LA INTELIGENCIA ARTIFICIAL

FABIO MORANDÍN-AHUERMA

PUEBLA Un gobierno presente
Secretaría de Educación
CONCYTEP Consejo de Ciencia y Tecnología del Estado de Puebla
Benemérita Universidad Autónoma de Puebla 1578
Complejo Regional Nororiental

PRINCIPIOS NORMATIVOS PARA UNA ÉTICA DE LA INTELIGENCIA ARTIFICIAL

Fabio Morandín-Ahuerma

Profesor Investigador de Tiempo Completo
Benemérita Universidad Autónoma de Puebla – Complejo Regional Nororiental, Arias y Boulevard s/n,
Col. del Carmen, Teziutlán, Puebla, C.P. 73800
fabio.morandin@correo.buap.mx

Agradecimientos

El autor agradece a todos quienes hicieron posible, de manera directa e indirecta, este proyecto editorial. Especialmente al Dr. Victoriano Covarrubias Salvatori, director general del Consejo de Ciencia y Tecnología del Estado de Puebla (CONCYTEP) por su invaluable apoyo y la Presentación. A la Dra. María Elena Álvarez-Buylla Roces, directora del Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCYT), a través del Sistema Nacional de Investigadores (SNI). A la Dra. Lilia Cedillo Ramírez, rectora de la Benemérita Universidad Autónoma de Puebla (BUAP) y al Dr. Sergio Díaz Carranza, director del Complejo Regional Nororiental por las facilidades otorgadas para la realización de este trabajo.

A los miembros del *AI Ethics Lab* de EIILM University, Dr. Ram Devi y Dra. Sunita Kumar por sus valiosos comentarios al primer borrador. Lo mismo al Dr. João Carlos Macedo, de la Universidade do Minho. Al Dr. en Ciencias, Fabian Valera Rivera, por su asesoría técnica. Al Mtro. Luis Gerardo Aguirre Rodríguez, de CONCYTEP, por el esmerado proceso editorial. A la Lic. Frida Tenorio y al Lic. Eduardo Jáuregui, por su atinada corrección de estilo. A la Lic. Monserrat Gómez e Iván Tovilla del curso de metodología de la BUAP por la pre-lectura, así como al Mtro. Victor Hugo García Castañon, de la UAMP.

A los revisores externos en el proceso anónimo par ciego por sus observaciones críticas y recomendaciones generales y específicas para la mejora de la versión final. A mis colegas Dr. Abelardo Romero Fernández y Mtra. Laura Villanueva Méndez, del Cuerpo Académico BUAP-CA-354 por su acompañamiento.

A mi familia, por su amor y apoyo incondicionales.

Sergio Salomón Céspedes Peregrina
Gobernador Constitucional del Estado de Puebla

Javier Aquino Limón
Secretario de Gobernación del Estado de Puebla

Gabriela Bonilla Parada
Presidenta del Sistema Estatal para el
Desarrollo Integral de la Familia

María Isabel Merlo Talavera
Secretaria de Educación del Estado de Puebla

Eduardo Castillo López
Presidente de la Junta de Gobierno y Coordinación Política del
H. Congreso del Estado Libre y Soberano de Puebla

Margarita Gayosso Ponce
Presidenta del Tribunal Superior de Justicia del Estado de Puebla

Victoriano Gabriel Covarrubias Salvatori
Director General del Consejo de Ciencia y Tecnología
del Estado de Puebla

Luis Gerardo Aguirre Rodríguez
Responsable del Área de Publicaciones

Frida Tenorio Espinosa
Eduardo Jáuregui Sainz de Rozas
Corrección de estilo

Luis Gerardo Aguirre Rodríguez
Diseño editorial

Primera edición, México, 2023

Publicado por el Consejo de Ciencia y Tecnología de Puebla
(CONCYTEP) B Poniente de La 16 de Sept. 4511,
Col. Huexotitla, 72534. Puebla, Pue.

ISBN: 978-607-8901-78-4

CÓDIGO IDENTIFICADOR CONCYTEP: C-L-2023-09-131

Esta obra para ser publicada fue dictaminada bajo la modalidad de pares a doble ciego por expertos en la materia.

Fabio Morandín-Ahuerma
Autor

Índice

Presentación

Como producto del esfuerzo coordinado entre el Gobierno del Estado de Puebla, que encabeza el Lic. Sergio Salomón Céspedes Peregrina, Gobernador Constitucional, y el Consejo de Ciencia y Tecnología del Estado de Puebla, que me honro presidir, nuestra misión ha sido atender el mayor número de líneas de investigación propuestas por las propias Dependencias del Ejecutivo a través de investigación básica y aplicada en Humanidades, Ciencia, Tecnología e Innovación, vinculando los sectores público, productivo, académico, ambiental y social. Las investigaciones publicadas se han convertido no solo en sugerencias, sino en insumos de políticas públicas y acciones específicas de este Gobierno.

Por lo anterior, dentro del Plan Estatal de Desarrollo 2019-2024, se encuentran la educación de calidad, el combate a la pobreza, el desarrollo económico para todas y todos, y la disminución de las desigualdades. En este sentido, la I+D (investigación y desarrollo) ha sido un tema recurrente, especialmente en este año con la aparición de nuevas tecnologías digitales. La construcción de un marco normativo para la inteligencia artificial puede generar oportunidades de desarrollo y acceso a la información, pero también implica riesgos para los que debemos estar preparados y capacitados. Esperamos que este libro sea una guía, tanto para los especialistas como para el público en general, interesados en las tecnologías que hoy recorren nuevos caminos de frontera de la ciencia.

Puebla, Puebla

Dr. Victoriano Gabriel Covarrubias Salvatori

Director General del Consejo de Ciencia y Tecnología del Estado de Puebla

Preámbulo

En el último lustro se han publicado múltiples documentos avalados por diferentes organizaciones, pero con un solo propósito: crear un marco de principios éticos cuya observancia garantice el desarrollo de una inteligencia artificial (IA) en beneficio de los usuarios y, en general, de la sociedad. Sin embargo, son tantos los principios que se han propuesto que muchas veces se genera confusión y, en el peor de los casos, es imposible su sistematización y aplicación porque algunos de estos principios son ambiguos y polisémicos; esto es que tienen varios significados e incluso pueden llegar a ser contradictorios.

El propósito de este trabajo es revisar esa lista, analizarla y tratar de proponer los principios éticos fundamentales que deberían observar los desarrolladores de software al programar algoritmos y, sobre todo, rescatar de la ética clásica algunos elementos para facilitar la comprensión del tema, descomponer la discusión estéril de algunos conceptos y resumirlos en honestidad, intencionalidad y conciencia moral.

En un mundo donde la inteligencia artificial se ha convertido en una parte integral de la vida diaria de millones de personas, los principios y las consideraciones éticas que rodean su desarrollo y uso nunca fueron tan importantes.

Una buena definición de inteligencia artificial redactada en 2023 por el Grupo de Trabajo Nacional de Recursos de Investigación de IA de los Estados Unidos (NAIRR) se refiere a que IA es un sistema basado en una máquina que puede hacer predicciones, recomendaciones o tomar decisiones que influyan en entornos reales o virtuales, para un conjunto dado de objetivos definidos por el ser humano. Los sistemas de inteligencia artificial utilizan entradas basadas en máquinas y humanos para percibir escenarios reales y virtuales; abstraer dichas percepciones en modelos mediante análisis de forma automatizada, y utilizar la inferencia de modelos para formular opciones de información o acción.

Actualmente, se tiene una explosión de inteligencia artificial gracias a la liberación de los modelos de lenguaje generativo. Lo mismo los creadores de texto a imagen y modelos híbridos multimodales. Otras muestras asombrosas de IA son aquellas capaces de dar un posible diagnóstico médico en segundos. De hecho, los propios involucrados en el desarrollo de IA están pidiendo una suspensión a su desarrollo, en tanto no se conozca con certeza las consecuencias positivas y negativas

venideras. De ahí la urgencia de conocer por todos estos principios normativos para una ética de la inteligencia artificial, que es el título de este libro.

Para ello, los "Principios de Asilomar" del Instituto del Futuro de la Vida son un conjunto de pautas desarrolladas por expertos en el campo, quienes describen principios clave para una IA segura y beneficiosa. Esta propuesta se une a la "Declaración de Montreal por una IA responsable" (La Déclaration de Montréal pour un développement responsable de l'intelligence artificielle 2018), que enfatiza la importancia del desarrollo y despliegue de la tecnología en un marco social ético.

En Europa también ha surgido un enfoque único de la ética de la IA, centrado en los derechos humanos y la transparencia. Las "Directrices éticas para una IA fiable" (Ethics guidelines for trustworthy AI) enfatizan la necesidad de que la IA se alinee con los valores y aspiraciones universales que protegen los derechos individuales.

A escala global, la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) también ha emitido recomendaciones sobre la ética de la IA. Esta "Recomendación sobre la ética de la inteligencia artificial adoptada el 23 de noviembre de 2021 [SHS/BIO/REC-AIETHICS/2021]" destaca la importancia de garantizar que la IA promueva la dignidad humana, la privacidad y la diversidad.

Del mismo modo, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) ha redactado un conjunto de principios para el desarrollo y despliegue responsable de la IA denominado "Recomendación del consejo de la OCDE sobre inteligencia artificial [OECD-LEGAL-0449]". Estos principios reconocen la transparencia, la rendición de cuentas y la inclusión.

Por su parte, desde la sociedad civil, la "Asociación sobre IA en beneficio de las personas y la sociedad" (Partnership on AI – PAI) es una colaboración entre los principales actores de la industria tecnológica, organizaciones sin fines de lucro e instituciones académicas para garantizar que la IA se desarrolle y utilice de una manera que beneficie a las personas y a la sociedad y ha emitido una serie de documentos en los que se especifica una postura ética proactiva hacia la IA.

Tampoco se deben minimizar los esfuerzos por construir un marco ético normativo que las propias empresas que desarrollan tecnología de punta como Google, Microsoft, Meta y Apple, entre otras, están construyendo para beneficio y protección de sus usuarios y de sus inversores en materia de algoritmos de inteligencia artificial, así como otras políticas empresariales autoimpuestas.

Los países líderes en el desarrollo de la inteligencia artificial, como lo son Estados Unidos (National AI Initiative Act of 2020 – Iniciativa Nacional de IA), China (新一代人工智能伦理规范 - Código de ética de la inteligencia artificial de nueva generación) y Rusia (Кодекс этики в сфере ИИ - Código de Ética de IA), son solo la muestra de que también las naciones están redactando sus respectivas políticas de Estado para el desarrollo y aplicación de una IA ética y benéfica para uso civil.

Juntos, estos diversos principios y recomendaciones brindan una hoja de ruta para el desarrollo y despliegue responsable de IA, asegurando que el potencial de esta tecnología transformadora se realice de una manera que ayude a toda la humanidad.

Sin embargo, al final de esta larga revisión de la literatura de los principios actuales, se ofrece una lectura crítica a modo de conclusión en el capítulo “Menos, es más: reconstruir una ética clásica normativa para un futuro responsable de la IA”.

Muchos filósofos, científicos y pensadores han reflexionado sobre estos temas a lo largo de los años y han contribuido al debate sobre la ética aplicada a la inteligencia artificial, por eso el interés de explicitar aquí los fundamentos para el desarrollo de una IA que sirva a la humanidad para cumplir sus más preciados anhelos y ponderar cuáles podrían ser los riesgos acuciantes que deben solventarse.

Se espera que la lectura de este trabajo sea fructífera sin importar el nivel de conocimiento que se tenga sobre el tema. El Consejo de Ciencia y Tecnología del Estado de Puebla se ha convertido en un divulgador nato de los avances de la ciencia, de sus ventajas y retos, que, al mismo tiempo, son los objetivos compartidos que se persiguen con la publicación de este trabajo de investigación.

VEINTITRÉS PRINCIPIOS DE ASILOMAR PARA LA INTELIGENCIA ARTIFICIAL Y EL FUTURO DE LA VIDA

Introducción

La “Conferencia de Asilomar, California sobre IA beneficiosa” fue una conferencia organizada por el Instituto del futuro de la vida (Future of Life Institute) en enero de 2017, donde más de cien expertos e investigadores se reunieron para discutir y formular principios para una IA ética. Los veintitrés principios están divididos en temas o preguntas de investigación (cinco); temas concernientes a la ética y valores (cinco) y, problemas a largo plazo (cinco). En este capítulo se abordan cada uno de los principios, se explicitan y, finalmente, se analiza y discute su viabilidad y vigencia. Se concluye que el movimiento ha recibido tanto aclamaciones como críticas. Mientras que algunos han elogiado los principios como un valioso punto de referencia para los debates sobre la ética de la IA, otros han expresado su preocupación de que carecen de orientación específica sobre cómo aplicarlos en la práctica. Se plantea la pregunta de si estos principios son suficientes para abordar los complejos retos éticos que surgen con la inteligencia artificial. La respuesta es que, aunque podrían no ser suficientes por sí solos, son necesarios para iniciar y dar forma a este debate.

Los principios de Asilomar de la IA

Los llamados “Principios de inteligencia artificial de Asilomar” (Asilomar AI Principles) [1] fueron desarrollados en el marco de una conferencia que tuvo lugar en Asilomar, California, en enero de 2017 y que fue organizada por el Instituto del futuro de la vida (Future of Life Institute) una organización filantrópica que se dedica a desarrollar iniciativas de solución a los problemas más urgentes del mundo, entre los que se encuentra la inteligencia artificial: “Desde algoritmos de recomendación hasta automóviles autónomos, la IA está cambiando nuestras vidas. A medida que aumenta el impacto de esta tecnología, también aumentan sus riesgos” [2, p. 1], advierte el Instituto.

Si bien la conferencia no estuvo abierta al público, entre los asistentes se encontraban Elon Musk, Francesca Rossi, Nick Bostrom, Peter Norvig, Ray Kurzweil, Sang Yong Lee, Stephen Hawking, Stuart Russell, Yann LeCun, Yoshua Bengio, entre otras personalidades. La conferencia también produjo una serie de catorce videos titulada “Beneficial AI 2017” (IA beneficiosa 2017) que están en YouTube [3].

Los acuerdos finales de la conferencia incluyen principios que abarcan una amplia gama de temas, como la transparencia, la responsabilidad, la privacidad y la seguridad. También contemplan la responsabilidad de los científicos y profesionales de la tecnología por el impacto de sus investigaciones y la necesidad de involucrar a una amplia gama de participantes en la toma de decisiones sobre el desarrollo y el uso de la IA.

La conferencia reunió a más de cien expertos de todo el mundo de diferentes disciplinas para discutir cómo asegurar que el desarrollo y uso de la IA beneficien a la sociedad y no representen un riesgo para los seres humanos. Al final de la conferencia, el 6 de enero de 2017, los participantes acordaron un conjunto de principios éticos y de responsabilidad para guiar el desarrollo y aplicación de sistemas de IA [1].

La “Conferencia de Asilomar, California sobre IA beneficiosa” de 2017 fue la continuación de una conferencia anterior celebrada en 2015 en Puerto Rico denominada “El futuro de la IA: oportunidades y retos” [5]. La conferencia de 2015 también fue organizada por el Instituto del Futuro de la Vida y se centró en debatir los posibles riesgos y beneficios de la inteligencia artificial. La conferencia de 2017 se basó en los debates y principios formulados en la conferencia anterior, pero se actualizó y se llegó a la formulación de veintitrés principios para una IA beneficiosa.

Los “Principios de Asilomar”, hasta hoy, son una importante guía para la comunidad de desarrolladores y han sido ampliamente difundidos, discutidos y analizados en el ámbito académico y en el mundo empresarial [6].

Se dividen en tres partes: la primera, investigación (cinco principios); la segunda, ética y valores (trece principios); y la tercera, aspectos en el largo plazo (cinco principios).

Temas de investigación

1. La investigación debe ser beneficiosa

> El objetivo de la investigación en IA no debe ser crear inteligencia no dirigida, sino inteligencia beneficiosa
> *[1, p. 1]*

Los investigadores de IA deben esforzarse por crear sistemas de inteligencia artificial que tengan un impacto positivo en la sociedad y en el mundo en general. En lugar de simplemente desarrollar sistemas inteligentes sin ningún propósito o dirección específica, los investigadores deben trabajar para garantizar que sus desarrollos estén diseñados y destinados a mejorar la calidad de vida, la productividad y el bienestar humano en general.

Además, las consideraciones éticas deben estar en primera línea de la investigación y el desarrollo de sistemas. A medida que la tecnología de IA avanza, es esencial abordar preocupaciones como la privacidad de los datos y la imparcialidad. Los investigadores deben estar atentos para reconocer y mitigar los posibles daños causados, por lo que es importante crear una cultura de desarrollo responsable en la que se valoren y defiendan la transparencia y la rendición de cuentas. También, la colaboración interdisciplinar es crucial en la investigación, ya que permite una comprensión completa del impacto de la IA en la sociedad. Al dar prioridad a las consideraciones éticas y colaborar en distintos campos, los investigadores pueden desarrollar tecnologías que no solo hagan avanzar el *corpus* de conocimiento, sino también beneficien a la sociedad de manera concreta.

Hay que resaltar, en este sentido, el trabajo desde la academia que actualmente realizan, por ejemplo, el Instituto de Ética en IA de la Universidad de Oxford (The Ethics in AI Institute) [7]; el Instituto de Internet de Oxford (Oxford Internet Institute) [8]; el Programa inteligencia artificial centrada en el ser humano de la Universidad de Stanford (Stanford University Human-Centered Artificial Intelligence) [9] y, en México, la Sociedad Mexicana de Inteligencia Artificial (SMIA) [10], entre muchos otros organismos e instituciones de investigación en el mundo.

2. La investigación debe ser financiada

> Fondos de investigación: Las inversiones en IA deben ir acompañadas de financiación para la investigación que garantice su uso beneficioso, incluyendo cuestiones polémicas en informática, economía, derecho, ética y estudios sociales *[1, p. 2]*.

Aquí se plantean, de acuerdo con el segundo principio de Asilomar, cuatro dilemas:

El primero se refiere a la necesidad de hacer que los futuros sistemas de IA sean robustos y confiables, de manera que funcionen sin fallas o vulnerabilidades a ataques cibernéticos. Esto es importante para garantizar que los sistemas puedan cumplir con sus funciones previstas sin causar daño o enviar errores que puedan tener consecuencias negativas.

El segundo dilema se centra en cómo se puede lograr un crecimiento económico a través de la automatización, sin dejar de lado los recursos y el propósito de las personas. Esto se refiere a la necesidad de garantizar que la IA no reemplace por completo los trabajos humanos y que se implemente de manera que mejore el bienestar y la calidad de vida de las personas, no que las deje sin trabajo.

El tercer dilema hace énfasis en la importancia de actualizar los sistemas legales para hacer frente a los riesgos asociados con la IA, al mismo tiempo que se mantiene la equidad y eficiencia en el proceso. Esto significa que los sistemas legales deben ser revisados y adaptados a la rápida evolución de la IA y sus implicaciones a corto, mediano y largo plazo.

La última cuestión aborda el tema de los valores éticos y legales que deben guiar el desarrollo y la implementación de la IA. Esto implica establecer un marco de gobernanza en concordancia con los valores humanos [1].

3. Vincular la ciencia con la política

> Enlace ciencia-política: Debe haber un intercambio constructivo y saludable entre los investigadores de IA y los actores políticos *[1, p. 3].*

Este principio reconoce la importancia de la colaboración entre la comunidad científica y los responsables gubernamentales a la hora de abordar los retos y oportunidades que presenta la inteligencia artificial. Reconoce que la IA tiene el potencial de transformar la sociedad de manera significativa y que su desarrollo y despliegue deben guiarse por consideraciones éticas y sociales.

El principio implica que los investigadores y los responsables políticos deben entablar un diálogo permanente para compartir información, preocupaciones y perspectivas sobre el desarrollo y el uso de la IA. Los investigadores pueden y deben informar a los políticos sobre los últimos avances en la tecnología y sus posibles aplicaciones, así como sobre los riesgos y retos que conllevan. Los responsables políticos pueden dar su opinión sobre las implicaciones legales, éticas y sociales de la IA y ayudar a configurar marcos reguladores y políticas públicas que reflejen los valores e intereses de la sociedad [11].

Algunos de los principales riesgos asociados a la IA son la pérdida de empleos debido a la automatización, la manipulación social a través de algoritmos (ver glosario) y construcción de perfiles, noticias falsas, vigilancia social a través de dispositivos, sesgos algorítmicos, desigualdades sociales, debilitamiento de los valores, armas autónomas, y algoritmos especulativos en los mercados de valores, entre otras muchas amenazas [44].

4. Generar una cultura de la investigación

> Cultura de investigación: Debe fomentarse una cultura de cooperación, confianza y transparencia entre investigadores y desarrolladores de IA *[1, p. 4].*

Este principio reconoce que el desarrollo e implantación de la IA es un esfuerzo de colaboración en el que participan investigadores, desarrolladores y otras partes interesadas. Asimismo, implica que deben trabajar juntos en un espíritu de

apertura, cooperación y respeto mutuo para avanzar en el desarrollo de la IA de forma responsable y beneficiosa.

El desarrollador y el investigador son dos roles laborales relacionados pero distintos en el campo de la inteligencia artificial. Un desarrollador es responsable de diseñar, perfeccionar e implementar sistemas basados en IA utilizando lenguajes y marcos de programación. Un investigador, por otro lado, se centra en el avance de los aspectos teóricos y prácticos de la IA mediante la realización de experimentos, análisis de los nuevos algoritmos y la publicación de artículos [43].

El principio contempla que una cultura de la investigación que fomente la cooperación, la confianza y la transparencia puede contribuir a garantizar que el desarrollo de la IA esté en consonancia con los valores y objetivos colectivos. Investigadores y desarrolladores pueden trabajar juntos para compartir conocimientos, experiencia y recursos para aplicar mejores prácticas y directrices éticas.

Además, este principio subraya la importancia de la transparencia en el desarrollo de la IA. Esto significa que los desarrolladores deben ser abiertos sobre sus métodos, datos y conclusiones, y tratar de entablar un diálogo con otras partes interesadas, incluidos los responsables políticos, las organizaciones de la sociedad civil y el público en general. Al promoverse la transparencia, se genera confianza y se garantiza que el desarrollo tecnológico sea responsable [12].

5. Seguridad por encima de competitividad

> Evitar carreras [comerciales]: Los equipos que desarrollan sistemas de inteligencia artificial deben cooperar para evitar la disminución de las normas de seguridad *[1, p. 5].*

Este principio reconoce que el desarrollo de la IA es un campo demasiado competitivo y que puede haber incentivos para que los desarrolladores den prioridad a la velocidad y la eficiencia, por encima de la seguridad y las consideraciones éticas. El principio implica que se debe trabajar para establecer y cumplir las normas de seguridad, y se debe evitar tomar atajos para obtener ventajas competitivas y comparativas para salir al mercado antes que los demás.

El principio de evitar una carrera comercial es para no poner en riesgo los asuntos en materia de protección. Por el contrario, los desarrolladores deben dar prioridad a la seguridad y a las consideraciones éticas, y trabajar en colaboración

para establecer y cumplir las normas más elevadas a favor de sus usuarios y de la sociedad en general. Esto puede ayudar a que la IA no plantee riesgos o daños innecesarios por el afán de lucro.

Además, el principio de evitar una carrera sugiere que los desarrolladores no deben considerar las normas de seguridad como una carga o un obstáculo para la innovación, sino como un componente esencial de la creatividad responsable. Trabajando en colaboración para establecer y cumplir las normas, los desarrolladores pueden contribuir a fomentar la confianza pública en la IA y garantizar que su desarrollo y despliegue sea responsable y benéfico [13].

Ética y valores

6. La IA debe ser segura

> Seguridad: los sistemas de IA deben ser seguros y protegidos durante toda su vida operativa, y de manera verificable cuando corresponda y sea factible *[1, p. 6].*

Este principio reconoce que los sistemas de IA tienen el potencial de plantear riesgos y daños si no se diseñan, desarrollan y despliegan de forma responsable y segura.

El principio implica que la seguridad debe ser primordial a lo largo de todo el ciclo de vida, desde el diseño y el desarrollo, hasta el despliegue y el funcionamiento de la IA. Esto significa implantar características y mecanismos que garanticen que el sistema sea confiable en todos los contextos y situaciones hipotéticas [13].

También subraya la importancia de la verificabilidad, lo que significa que las afirmaciones de seguridad hechas por los desarrolladores deben poder probarse a través de medios independientes e incluso ataques controlados. Esto puede ayudar a garantizar que los sistemas de IA sean robustos en la práctica y no solo teóricamente.

7. Transparente en cuanto a sus errores

> Transparencia de fallas: Si un sistema de IA causa daños, debe ser posible determinar la causa *[1, p. 7].*

Este principio reconoce que los sistemas de IA no son infalibles y que pueden fallar e incluso causar perjuicios. El principio implica que, cuando un sistema de IA falla, es importante entender por qué y en dónde se equivocó, para evitar que se produzcan incidentes similares en el futuro, y jamás minimizar u ocultar lo sucedido [14].

El principio de transparencia en los fallos implica que los desarrolladores deben ser abiertos sobre cómo funciona el sistema, qué datos utiliza y cómo toma decisiones, evitando los denominados algoritmos de caja negra porque se desconoce qué sucede en su interior. Un algoritmo se compone básicamente de datos de entrada, proceso y salida. En términos sencillos, la relación entrada-salida de un algoritmo de caja negra es conocida, pero los pasos reales que sigue el algoritmo para llegar al resultado no son transparentes. Esta carencia puede dificultar la comprensión de cómo toma decisiones y por qué produce determinados resultados. Puede ser problemático en situaciones en las que las decisiones tomadas por el algoritmo tienen un impacto significativo en los individuos o en la sociedad [15]. Algunos ejemplos de algoritmos de caja negra son las redes neuronales, las máquinas de vectores de soporte (SVM por sus siglas en inglés) y algunos árboles de decisión automatizados, entre otros. Las SVM son un tipo de algoritmo de aprendizaje automático o *machine learning* (véase el glosario) que puede utilizarse para tareas de clasificación o regresión de los datos con etiquetas o valores y se puede aplicar a problemas como el análisis de señales, la comprensión artificial de lenguajes naturales y la identificación de imágenes y sonidos [16].

Además, cuando un sistema falla, es importante contar con procesos para comprender y documentar el fallo, incluyendo qué salió mal, por qué ocurrió y qué medidas pueden tomarse para evitar hechos similares en el futuro. En las cajas negras, esto no es posible [14].

Asimismo, la transparencia implica que se dé prioridad al aprendizaje, en lugar de culpar o encubrir errores. Al ser transparentes, los desarrolladores pueden ayudar a generar confianza pública en la IA y saber que, si algo sale mal, por lo menos se sabrá y, segundo, se buscará cómo rectificarlo.

8. Transparencia en asuntos judiciales

> Transparencia judicial: Cualquier participación de un sistema autónomo en la toma de decisiones judiciales debe proporcionar una explicación satisfactoria auditable por una autoridad humana competente *[1, p. 8].*

Este principio reconoce que los sistemas de IA se utilizan cada vez más en algunos sistemas judiciales para la evaluación de riesgos, recomendación de penas y otros procesos de toma de decisiones. El principio implica que, cuando un sistema de IA participa en la toma de decisiones judiciales es importante garantizar que dicha decisión pueda explicarse de forma simple, comprensible y verificable por una autoridad humana [17]. Por ejemplo, el Tribunal Popular Intermedio de Hangzhou en China está utilizando IA para ayudar a los jueces a tomar decisiones. El sistema de IA, llamado Xiao Zhi 3.0, se ha utilizado en más de 10 000 casos hasta el momento [45].

El principio de transparencia judicial subraya la importancia de la responsabilidad y la explicabilidad en el uso de la IA en el sistema penal. Esto significa que los desarrolladores y usuarios de IA en los juzgados deben dar prioridad a proporcionar a los acusados explicaciones claras y comprensibles de cómo el sistema ha llegado a una determinación, toda vez que tiene implicaciones de largo alcance. Además, el principio hace énfasis en la importancia de la supervisión y participación humanas en el proceso de toma de decisiones para garantizar que sean coherentes con las normas legales y éticas, y que puedan ser revisadas en todo momento [17].

La transparencia judicial implica que los sistemas de IA utilizados deben diseñarse y desarrollarse teniendo en cuenta la rendición de cuentas en caso de apelaciones [14].

9. Responsabilidad

> Los diseñadores y usuarios de sistemas avanzados no pueden minimizar la responsabilidad humana cuando las decisiones han sido tomadas por la IA *[1, p. 9].*

Este principio reconoce que el desarrollo y el despliegue de IA tiene implicaciones morales, y que quienes diseñan y construyen estos sistemas tienen la

responsabilidad de considerar y dar forma a los alcances que estos tengan. El principio añade que los diseñadores y constructores de sistemas de IA deben desempeñar un papel proactivo en la configuración de las implicaciones morales de sus propios sistemas, en lugar de limitarse a reaccionar ante ellos como un agente pasivo [19].

El principio de responsabilidad subraya la importancia de las consideraciones éticas en el diseño, desarrollo y despliegue de los sistemas de IA. Esto significa que los diseñadores y constructores deben dar prioridad a la consideración de los impactos y consecuencias potenciales de sus sistemas, y tomar medidas para garantizar que éstos se ajusten a normas éticas y morales [19] y en caso de que no sea así, responder diligentemente por ello.

Por ejemplo, si se hace uso de código generado por IA esto es especialmente arriesgado si los usuarios no pueden validarlo, ya sea porque no tienen suficiente conocimiento técnico o porque la herramienta disuade a los usuarios de verificar su salida. Para mitigar estos riesgos, se debe tratar el código generado por IA de la misma manera que se trataría su contraparte escrita por humanos. Eso significa aplicar las mismas políticas de seguridad y responsabilidad en todos los ámbitos, ya sea que el código de programación provenga de un ser humano o de un modelo de IA [46].

10. Alineación de valores

> Valores alineados: Los sistemas de IA altamente autónomos deben diseñarse de modo que sus objetivos y comportamientos puedan alinearse con valores humanos a lo largo de su operación *[1, p. 10].*

Este principio reconoce los riesgos potenciales asociados a los sistemas de IA autónomos, capaces de tomar decisiones y emprender acciones sin intervención humana. El principio implica garantizar que estos sistemas estén diseñados de forma que se alineen con valores humanos, para evitar que actúen de forma perjudicial o incoherente con las normas establecidas [21].

Esto significa que los diseñadores y constructores de sistemas de IA deben dar prioridad a garantizar que los objetivos y comportamientos de los sistemas no actúen al margen de la ley, y que no lo hagan de forma perjudicial violando la

primera regla de Asimov: “Una máquina no puede dañar a un ser humano, o por inacción permitir que un ser humano sufra daño” [22].

11. Respeto por los derechos humanos

> Valores humanos: Los sistemas de IA deben diseñarse y funcionar de modo que sean compatibles con los ideales de dignidad humana, derechos, libertades y diversidad cultural *[1, p. 11].*

Este principio implica que los sistemas de IA no solo deben ajustarse a normas éticas y morales, sino también a valores más amplios relacionados con la dignidad, los derechos humanos inalienables, la libertad y la diversidad cultural.

El principio de los valores humanos subraya la importancia de diseñar y hacer funcionar los sistemas de IA de forma que respeten el valor intrínseco de la persona. Esto significa que los diseñadores y constructores de sistemas deben dar prioridad a garantizar que no atenten contra la dignidad humana, vulneren los derechos humanos o limiten la diversidad cultural [23].

Las empresas que utilizan software de IA para tomar decisiones sobre salud y medicina, empleo e incluso justicia penal, por ejemplo, deben responder cómo se aseguran de que los programas no estén codificados, consciente o inconscientemente, con sesgos estructurales. La adopción más amplia de la IA en la atención médica, los vehículos autónomos y en otras industrias depende del marco que determina quién, si es que alguien, termina siendo responsable de una violación a los derechos básicos de las personas por los sistemas de IA [24].

12. Privacidad de la información

> Privacidad personal: Las personas deben tener derecho a acceder, gestionar y controlar los datos que generan, dado el poder de los sistemas de IA, para analizar y utilizar esos datos *[1, p. 12].*

Este principio reconoce la importancia cada vez mayor de los datos personales en la era de la IA, y los riesgos potenciales asociados al mal uso o al manejo inadecuado de información personal. Implica que las personas deben tener el control de los datos que generan y el derecho a borrarlos, así como acceder a ellos de forma coherente con sus propios intereses y preferencias.

La lista de datos personales incluye: nombre y apellido, cónyuge, hijos, familia, dirección física y dirección de correo electrónico, teléfono, fecha de nacimiento, género, nacionalidad, pasaporte o número de identificación oficial, información financiera, tarjetas de crédito o detalles de cuentas bancarias, número de seguro social, RFC o cualquier registro como contribuyente, CURP, DNI o número similar en cada país, información médica, biométrica, empleo, número de personal, educación, calificaciones, dirección IP, geolocalización, nombres de usuario, entre otros.

El principio de privacidad subraya la importancia de proteger los datos personales y no compartir información que no es de la incumbencia de empresas, particulares e incluso gobiernos. Esto significa que los diseñadores y constructores de sistemas de IA deben dar prioridad al diseño de sistemas que sean transparentes y responsables con respecto a la gestión y el uso de los datos, y que permitan a las personas ejercer el control sobre su propia información.

Además, el principio de privacidad personal implica que estos datos deben estar protegidos contra el acceso no autorizado o el uso indebido. Es sabido que particulares sin escrúpulos venden grandes bases de datos en el mercado negro de datos personales [25]. Por tanto, se debe dar prioridad al diseño de sistemas que respeten y protejan la privacidad personal y que permitan a los individuos ejercer el control sobre sus datos [26].

13. Compatibilidad entre privacidad y libertad

> Libertad y privacidad: La aplicación de la IA a los datos personales no debe restringir injustificadamente la libertad real o percibida de las personas *[1, p. 13].*

Este principio reconoce los riesgos potenciales asociados al uso de tecnologías de IA para analizar y procesar datos personales, y la necesidad de garantizar que dicho uso no conlleve infracciones injustificadas de las libertades individuales. El principio destaca la importancia de equilibrar los beneficios de la tecnología de IA con la necesidad de proteger las libertades individuales y la privacidad [27].

El principio de libertad y privacidad implica que la aplicación de la IA a los datos personales debe estar sujeta a medidas reguladoras adecuadas, con el fin de garantizar que las personas no vean indebidamente restringida su capacidad para ejercer sus derechos y libertades. Por ejemplo, los sistemas de IA no deben utilizarse para recabar datos biométricos de manera inadvertida para las personas,

pues ello es considerado una violación a su privacidad, léase, reconocimiento de rostro, grabación de voz, escaneo del iris o cualquier otro dato biométrico.

14. Beneficios para todos

> Beneficios compartidos: Las tecnologías de IA deben beneficiar y empoderar al mayor número de personas posible *[1, p. 14].*

El principio del beneficio compartido implica que las tecnologías de la IA no solo deben desarrollarse y desplegarse en beneficio de unos cuantos privilegiados, sino que deben ser accesibles y estar disponibles para todos. Esto significa que las tecnologías de la IA deben diseñarse teniendo en cuenta las necesidades y los intereses de diversas comunidades, y que los beneficios de estas tecnologías deben distribuirse de forma justa y equitativa, considerando todas las regiones del planeta, especialmente África y América Latina.

Además, el principio del beneficio compartido subraya la importancia de capacitar a las personas y las comunidades para puedan utilizar las tecnologías de la IA con el fin de alcanzar sus objetivos y aspiraciones.

15. Compartir la prosperidad

> Prosperidad compartida: La prosperidad económica creada por la IA debe compartirse ampliamente, en beneficio de toda la humanidad *[1, p. 15].*

El principio de prosperidad compartida implica que las tecnologías de la IA no solo deben beneficiar a los pocos que las controlan o poseen, sino a todos los miembros de la sociedad. Exige que las ganancias y beneficios económicos generados por las tecnologías de IA se compartan, de forma que beneficien a un amplio sector.

Los beneficios económicos de la IA no deben concentrarse en manos de unos pocos individuos o empresas, sino distribuirse de forma más equitativa entre la sociedad. Esto podría lograrse a través de políticas y programas que promuevan una mayor igualdad de ingresos y acceso a las oportunidades económicas, así como a través de iniciativas que apoyen el desarrollo y despliegue de las tecnologías de IA de manera que se democraticen [28].

Algunos sistemas de IA son alimentados por millones de trabajadores mal pagados en todo el mundo, quienes realizan tareas repetitivas en condiciones laborales precarias, lejos de Silicon Valley. Estos trabajadores explotados a menudo se reclutan entre países con mano de obra barata como Kenia, India, Filipinas e incluso México [47].

16. Control humano

> El ser humano debe decidir si delega o no sus decisiones en los sistemas de inteligencia artificial para alcanzar los objetivos que haya elegido *[1, p. 16].*

El principio de control humano reconoce que las tecnologías de IA pueden automatizar muchas tareas y procesos de toma de decisiones que antes realizaban solo los humanos. Sin embargo, también hace énfasis en que se debe conservar la capacidad de decidir cuándo y cómo delegar la toma de decisiones de esos sistemas. Esto implica que los desarrolladores deben tener la responsabilidad y la autoridad últimas para tomar decisiones y fijar objetivos para sus sistemas, basándose en valores predeterminados [29] [30].

Además, el principio de control humano también sugiere que debe tenerse la capacidad de anular o intervenir en los procesos de toma de decisiones automatizadas, si fuera necesario. Esto puede ser especialmente importante en situaciones en las que la IA pueda tener implicaciones éticas o morales significativas.

El principio de control humano también subraya la importancia de la transparencia en los procesos de toma de decisiones de los sistemas de IA. Esto significa que los seres humanos deben ser capaces de entender cómo los sistemas llegan a sus decisiones y deben tener acceso a la información sobre los algoritmos y los datos utilizados para fundamentar dichas decisiones. Antes se mencionó el riesgo de los llamados algoritmos de caja negra [31].

17. Evitar la disrupción

> No subversión: El poder que confiere el control de sistemas de IA muy avanzados debe respetar y mejorar, en lugar de trastornar, los procesos sociales y cívicos de los que depende la salud de la sociedad *[1, p. 17].*

El principio de no subversión se refiere a evitar que los sistemas de IA se utilicen para manipular o socavar los procesos democráticos o los derechos y libertades individuales. Esto incluye garantizar que los sistemas de IA no se utilicen para difundir desinformación o manipular la opinión pública, así como garantizar que no violen los derechos de privacidad o permitan nuevas formas de discriminación. Son notorios los casos en que se ha manipulado a la opinión pública a través de la generación de mensajes de bots en redes sociales para modificar las preferencias políticas, incitar a la xenofobia o influir en sus creencias [32]. Un bot –que es una abreviatura de la palabra robot– es un programa informático que realiza tareas automatizadas, repetitivas y predefinidas, y que suele imitar o sustituir el comportamiento de los usuarios humanos.

Además, el principio de no subversión reconoce la importancia de garantizar que los sistemas de IA se desarrollen de forma transparente y responsable. Para ello, la IA debe estar sujeta a supervisión y regulación, y su desarrollo debe guiarse por principios éticos y el bien común.

18. Armisticio de IA

> Carrera armamentista de IA: Debe evitarse una carrera armamentista de armas autónomas letales *[1, p. 18].*

El principio de Asilomar advierte los peligros inminentes de la invención, producción y uso de armas autónomas letales (Lethal Autonomous Weapons o LAWs) [1] [48] y se refiere a la preocupación de que esto pueda conducir a una carrera armamentista en la que los países u otras entidades compitan por desarrollar y desplegar sistemas de IA cada vez más avanzados que tomen decisiones sobre la vida. Una competición de este tipo podría conducir a una peligrosa escalada de los conflictos y a la proliferación de armas mortíferas que no estén bajo control humano.

Este principio subraya la importancia de garantizar que el desarrollo de las tecnologías de IA esté guiado por un compromiso con los valores éticos y humanitarios, y que los responsables políticos tomen medidas para impedir el desarrollo de armas autónomas que puedan causar daños a civiles o exacerbar conflictos. Especialmente que no sean violados los Convenios de Ginebra de 1949, destinados a evitar la barbarie en conflictos armados y que no se comentan crímenes de lesa humanidad [33].

También subraya la necesidad de cooperación y coordinación internacionales para hacer frente a los retos que plantea el rápido desarrollo de la IA y su uso en contextos militares, tal como se está viviendo en la ofensiva militar de la Federación Rusa a Ucrania desde febrero de 2022 [34] y en otros conflictos armados.

En el largo plazo

19. Restricción sobre futuras capacidades

> Precaución de capacidad: al no haber consenso, se debe evitar suposiciones fuertes con respecto a los límites superiores en las futuras capacidades de la IA *[1, p. 19].*

Este principio reconoce que actualmente no hay consenso entre los expertos sobre cuáles podrían ser los límites de las capacidades de la IA, y que existe el riesgo de hacer predicciones demasiado optimistas o pesimistas sobre el potencial de las tecnologías. Algunos han puesto como límite las capacidades autogenerativas de GPT4 (Generative Pre-trained Transformer) (ver glosario) antes de llegar a una inteligencia artificial general o fuerte (AGI, por sus siglas en inglés) que puede hacer todo lo que un humano hace o siente.

El principio sugiere que, dada la incertidumbre que rodea el desarrollo de la IA, se debería evitar hacer suposiciones tajantes sobre los alcances que pueda llegar a tener. En su lugar, se debe adoptar un enfoque prudente, reconociendo los riesgos y beneficios potenciales de la IA y trabajando para garantizar que los sistemas se desarrollen de forma que estén en consonancia con los valores y las prioridades humanas [35].

20. Importancia del futuro de la tierra

> Magnitud: La IA avanzada podría representar un cambio profundo en la historia de la vida en la Tierra, y debería planificarse y gestionarse con el cuidado y los recursos adecuados *[1, p. 20].*

Este principio subraya que se debe tomar en serio el desarrollo de sistemas avanzados de IA y abordarlo con cautela, reconociendo los riesgos y beneficios potenciales. Sugiere que se debe invertir recursos para comprender mejor las posibles repercusiones de los sistemas avanzados de IA y desarrollar estrategias y políticas que garanticen que se desarrollan y utilizan de forma responsable y beneficiosa. Esto incluye la participación de una amplia gama de partes interesadas, como expertos en IA, responsables políticos, filósofos especialistas en ética, y miembros del público usuario, en el desarrollo y la gobernanza de estos sistemas [36].

21. Peligros de la IA

> Riesgos: Los conflictos que plantean los sistemas de IA, especialmente los riesgos catastróficos o existenciales, deben estar sujetos a esfuerzos de planificación y mitigación acordes con el impacto esperado *[1, p. 21].*

El principio hace énfasis en la importancia de identificar y mitigar los riesgos que plantean los sistemas de IA. La declaración reconoce que la IA tiene el potencial de ayudar, pero también causar daños significativos a la humanidad. Por lo tanto, se exige una planificación proactiva y esfuerzos de mitigación para abordar estos riesgos de manera responsable. La declaración también advierte que los riesgos potenciales de los sistemas de IA no se comprenden totalmente y que es necesario seguir investigando para identificarlos y evaluarlos [37].

A medida que la tecnología de IA sigue evolucionando es necesario mantenerse alerta sobre los posibles riesgos y prepararse en consecuencia. El impacto de estos riesgos podría ser enorme y, como tal, los recursos asignados para contenerlos deben ser proporcionales.

22. Automejora recursiva

> Superación autónoma recursiva: Los sistemas de IA diseñados para auto-mejorarse recursivamente o auto-replicarse de manera que puedan conducir a un rápido aumento de la calidad o la cantidad, deben estar sujetos a estrictas medidas de seguridad y control *[1, p. 22].*

El principio de superación automática recursiva se refiere a la capacidad de algunos sistemas de IA para aprender y mejorarse a sí mismos con el tiempo; esto se conoce como *deep learning* o aprendizaje profundo [38] (véase el glosario). La autosuperación recursiva permite a un sistema de IA aprender continuamente y tomar decisiones a un ritmo cada vez más rápido. Aunque esto puede ser una característica valiosa para los sistemas de IA, también tiene el potencial de crear riesgos y consecuencias no deseadas.

Por ello, estos sistemas deben estar sujetos a estrictas medidas de seguridad y control, y minimizar así el riesgo de crecimiento incontrolado o de consecuencias imprevistas derivadas de la mejora automática. El sistema de IA debe diseñarse y supervisarse cuidadosamente para garantizar que funciona dentro de los parámetros de seguridad establecidos y no suponga una amenaza para la seguridad o el bienestar humanos. Esto ha dado lugar a muchas series de ficción [39], pero también es una realidad que la autonomía total puede quedar fuera del control de las personas con efectos indeseados. Así es un menester la cautela, el control y la planificación cuidadosa al desplegar sistemas de IA que tengan potencial para la auto-mejora recursiva [40].

23. IA para el beneficio común

> Bien común: La superinteligencia solo debe desarrollarse al servicio de ideales éticos ampliamente compartidos, y en beneficio de toda la humanidad y no de un Estado u organización *[1, p. 23].*

Este principio está relacionado con el desarrollo de la inteligencia artificial general (AGI), que es, como ya se dijo, una inteligencia que tiene la capacidad de realizar cualquier tarea intelectual que pueda hacer un ser humano [41]. La superinteligencia artificial (ASI) en cambio, sería capaz de superar en la disciplina que sea a cualquier ser humano en capacidades cognitivas, habilidades, destrezas,

competencias, etcétera. Ambas inteligencias, hasta ahora, siguen siendo solo una posibilidad teórica.

Por ello, el principio hace énfasis en que el desarrollo de la superinteligencia, en caso de que se logre, no debe estar impulsado por intereses individuales, sino que debe servir al bien común y a los valores éticos ampliamente compartidos entre las distintas sociedades. Implica que el desarrollo de la AGI debe guiarse por un consenso mundial sobre los principios éticos que deben regir su desarrollo y utilización. En cambio, mientras no haya un marco normativo que las sancione, pueden representar un peligro para la humanidad. Durante el foro World Government Summit en Dubai en marzo de 2023, Elon Musk afirmó: “Uno de los mayores riesgos para el futuro de la civilización es la IA, es positiva o negativa y tiene una gran, gran promesa, gran capacidad… [pero también] un gran peligro” [42].

Por eso este principio es importante porque el desarrollo de la superinteligencia tiene el potencial de afectar significativamente al mundo y podría tener consecuencias de gran alcance para todos. Si llegaran a existir la AGI o la ASI, deberán estar al servicio de ideales éticos universales.

Conclusiones parciales

Hasta aquí los veintitrés principios formulados al término de la Conferencia de Asilomar que, debe señalarse, algunos han quedado superados por la mano invisible del mercado y la feroz competencia hegemónica-política por el control de la IA. Esto es, una carrera bélica autónoma en pleno desarrollo, falta de transparencia de muchos algoritmos por el secreto industrial y sesgos raciales o de clase, especialmente en materia judicial, laboral y de seguridad pública, son visibles.

Sin embargo, los principios de Asilomar son un conjunto de directrices propuestas para regular el desarrollo y uso de la IA de forma responsable. Los “Principios de Asilomar para la IA” de 2017 pretenden garantizar que se desarrolle de forma segura y que beneficie a la sociedad.

Si bien estos principios son un punto de partida importante para abordar las cuestiones éticas y de seguridad que plantea la IA, es importante señalar que no resolverán todos los problemas y retos asociados a la misma. Son necesarios debates continuos, intercambios entre las partes interesadas, normativas y esfuerzos concertados para garantizar que la IA se desarrolle y utilice de forma responsable.

Una crítica a los veintitrés principios de Asilomar y al movimiento más amplio de ética de la IA es que son demasiado generales y carecen de orientaciones específicas sobre cómo aplicarlos en la práctica. Algunos como Garbowski [6] se preguntan si son suficientes para abordar los complejos retos éticos que plantea la IA, pero podría responderse que sí lo son; si bien no son suficientes, sí son necesarios y representan un marco fundacional que debe seguir desarrollándose y perfeccionando para proporcionar orientaciones más específicas sobre la aplicación práctica de los principios éticos. En general, los veintitrés principios han sido acogidos receptivamente por la comunidad y a menudo se citan como punto de referencia clave en los debates formales e informales sobre ética de la IA.

Referencias

[1] Future of Life Institute, "The Asilomar AI Principles," Futureoflife.org Acceso ene. 2023. [En línea] Disponible: https://futureoflife.org/open-letter/ai-principles/

[2] Future of Life Institute, "Cause Area Artificial Intelligence," Futureoflife.org Acceso ene. 2023. [En línea] Disponible: https://futureoflife.org/cause-area/artificial-intelligence/

[3] Future of Life Institute, "Beneficial AI 2017." (30 de enero de 2017). [Video en línea]. Disponible: https://bsu.buap.mx/b0e

[4] Future of Life Institute, "Steering transformative technology towards benefitting life and away from extreme large-scale risks," Futureoflife.org Acceso ene. 2023. [En línea] Disponible: https://futureoflife.org/

[5] Future of Life Institute, "The Future of AI: Opportunities and Challenges," Futureoflife.org Acceso ene. 2023. [En línea] Disponible: https://futureoflife.org/event/ai-safety-conference-in-puerto-rico/

[6] M. Garbowski, "A critical analysis of the Asilomar AI principles," *Zeszyty Naukowe*, vol. 115, pp. 45-55, 2017, https://bsu.buap.mx/cjb

[7] Oxford Institute for Ethics in AI, "Institute for Ethics in AI." Acceso ene. 2023. [En línea] Disponible: https://www.oxford-aiethics.ox.ac.uk/

[8] Oxford Internet Institute, "Oxford Internet Institute," Acceso ene. 2023. [En línea] Disponible: https://www.oii.ox.ac.uk/

[9] HAI. "Stanford University Human-Centered Artificial Intelligence," Stanford.edu. Acceso ene. 2023. [En línea] Disponible: https://hai.stanford.edu/

[10] SMIA, "Sociedad Mexicana de Inteligencia Artificial," Acceso ene. 2023. [En línea] Disponible: https://smia.mx/

[11] V. Durrer, T. Miller, L. A. Celi, y M. Ghassemi, "The Routledge Handbook of Global Cultural Policy," 1st ed. Abingdon: Routledge, 2018.

[12] J. Kroll, "Accountability in Computer Systems," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 180-196.

[13] C. Stadlmann y A. Zehetner, “Human Intelligence Versus Artificial Intelligence: A Comparison of Traditional and AI-Based Methods for Prospect Generation,” en *Marketing and Smart Technologies,* Springer, 2021, pp. 11-22.

[14] N. Diakopoulos, “Transparency. Accountability, Transparency, and Algorithms,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford: Oxford University Press, 2020, pp. 197-213, doi: 10.1093/oxfordhb/9780190067397.013.11.

[15] L. Floridi, “*Ethics, Governance, and Policies in Artificial Intelligence*,” Cham: Springer, 2021.

[16] G. Z. Yang et al., “The grand challenges of Science Robotics,” *Science Robotics*, vol. 3, no. 14, p. eaar7650, ene. 2018, doi: 10.1126/scirobotics.aar7650

[17] H. Surden, “Ethics of AI in Law: Basic Questions,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford: Oxford University Press, 2020, pp. 719-736.

[18] W. Schröder, “Robots and Rights: Reviewing Recent Positions in Legal Philosophy and Ethics,” en *Robotics, AI, and Humanity: Science, Ethics, and Policy*, J. von Braun et al., Eds., Springer International Publishing, Cham, 2021, pp. 191-203.

[19] C. Bartneck, C. Lütge, A. Wagner, y S. Welsh, “Responsibility and Liability in the Case of AI Systems,” en *An Introduction to Ethics in Robotics and AI*, C. Bartneck, et al., Eds. Springer International Publishing, 2021, pp. 39-44.

[20] D. Gunkel, “Perspectives on Ethics of AI: Philosophy,” in *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 538-553.

[21] A. Korinek, “Integrating Ethical Values and Economic Value to Steer Progress in Artificial Intelligence,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 475-491.

[22] I. Asimov, “Strange playfellow,” *Super Science Stories*, vol. 1, no. 4, pp. 67-77, 1940.

[23] L. Lim y H. K. Lee, “Routledge handbook of creative and cultural industries in Asia,” Routledge handbooks, Abingdon, UK: Routledge, 2019.

[24] K. Yeung, A. Howes, y G. Pogrebna, “AI Governance by Human Rights-Centered Design, Deliberation, and Oversight: An End to Ethics Washing,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 76-106.

[25] C. Jordan y D. Maimon, “New research shows that darknet markets net millions selling stolen personal data,” Fastcompany.com. Acceso ene. 2023. [En línea] Disponible: https://bsu.buap.mx/b0G

[26] J. Antoniou y O. Tringides, “Personal Data, Cloud Platforms, Privacy and Quality of Experience,” in *Effects of Data Overload on User Quality of Experience*, J. Antoniou y O. Tringides, Eds., Cham: Springer International Publishing, 2023, pp. 37-54.

[27] C. Bartneck, C. Lütge, A. Wagner y S. Welsh, “Privacy Issues of AI,” in *An Introduction to Ethics in Robotics and AI*, C. Bartneck et al., Eds., Springer, 2021, pp. 61-70.

[28] H. Dang y P.F. Lanjouw, "Toward a New Definition of Shared Prosperity: A Dynamic Perspective from Three Countries," en *Inequality and Growth: Patterns and Policy: Volume I: Concepts and Analysis*, K. Basu and J.E. Stiglitz, Eds., Palgrave, 2016, pp. 151-171.

[29] F. Morandín-Ahuerma, "Leyendas de trolley: juicio moral y toma de decisiones," *Universita Ciencia*, vol. 8, no. 22, pp. 79-91, 2019.

[30] J. Morley, L. Floridi, L. Kinsey y A. Elhalal, "From What to How: An Initial Review of Publicly Disponible en AI Ethics Tools, Methods and Research to Translate Principles into Practices," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer International Publishing, 2021, pp. 153-183.

[31] L. Ibarra, D. Balderas, P. Ponce y A. Molina, "Fast Execution of Black-Box Algorithms Through a Piece-Wise Linear Interpolation Technique," *Arab. J. Sci. Eng.*, vol. 44, no. 11, pp. 9443-9453, 2019, doi: 10.1007/s13369-019-04042-y.

[32] J. Mökande, J. Morley, M. Taddeo y L. Floridi, "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations," *Sci. Eng. Ethics*, vol. 27, no. 4, p. 44, 2021, doi: 10.1007/s11948-021-00319-4.

[33] J. Galliott y J. Scholz, "The Case for Ethical AI in the Military," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 684-702, doi: 10.1093/oxfordhb/9780190067397.013.43.

[34] S. Russell, "AI weapons: Russia's war in Ukraine shows why the world must enact a ban," *Nature*, vol. 614, no. 7949, pp. 620-623, 2023.

[35] T. Powers y J. Ganascia, "The Ethics of the Ethics of AI," in *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 26-51, doi: 10.1093/oxfordhb/9780190067397.013.2.

[36] S. Russell y P. Norvig, "Philosophy, ethics, and safety of AI," en *Artificial Intelligence: A Modern Approach*, Londres: Pearson, 2022, pp. 1032-1062.

[37] T. Winkle, "Product Development within Artificial Intelligence, Ethics and Legal Risk." Cham: Springer Vieweg.

[38] H.P. Cowley et al., "A framework for rigorous evaluation of human performance in human and machine learning comparison studies," *Scientific Reports*, vol. 12, no. 1, p. 5444, 2022, doi: 10.1038/s41598-022-08078-3.

[39] E. Saffari, S. R. Hosseini, A. Taheri y A. Meghdari, "Does cinema form the future of robotics? a survey on fictional robots in sci-fi movies," *SN Applied Sciences*, vol. 3, no. 6, p. 655, 2021, doi: 10.1007/s42452-021-04653-x.

[40] A. Majot y R. Yampolskiy, "Diminishing Returns and Recursive Self Improving Artificial Intelligence," en *The Technological Singularity: Managing the Journey*, V. Callaghan et al., Eds., Springer Berlin Heidelberg, 2017, pp. 141-152.

[41] T. Vasil, P. Skalfist, y D. Mikelsten, "Inteligencia artificial: la cuarta revolución industrial," Cambridge Stanford Books, 2020.

[42] T. Barrabi, “Elon Musk warns AI ‘one of biggest risks’ to civilization during ChatGPT’s rise,” NYPost.com, Acceso ene. 2023. [En línea] Disponible: https://nypost.com/2023/02/15/elon-musk-warns-ai-one-of-biggest-risks-to-civilization/

[43] S. Gupta, “Data Scientist vs. Artificial Intelligence Engineer: Which Is a Better Career Choice?”, Acceso ene. 2023. [En línea] Disponible: https://bsu.buap.mx/b3P

[44] M. Thomas, “8 Risks and Dangers of Artificial Intelligence to Know.” Builtin.com. Acceso ene. 2023. [En línea] Disponible: https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence.

[45] DeutscheWelle, “Cortes chinas ya resuelven casos con inteligencia artificial.” DW.com. Acceso ene. 2023. [En línea] Disponible: https://www.dw.com/es/las-cortes-de-china-ya-utilizan-inteligencia-artificial-para-resolver-casos/a-64471873

[46] L. Craig, “The promises and risks of AI in software development,” Techtarget, Acceso ene. 2023, [En línea]. Disponible: https://www.techtarget.com/searchitoperations/feature/The-promises-and-risks-of-AI-in-software-development

[47] A. Williams, M. Miceli y T. Gebru, “The Exploited Labor Behind Artificial Intelligence,” Acceso ene. 2023. [En línea]. Disponible: https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/

[48] M. Taddeo y A. Blanchard, “Accepting Moral Responsibility for the Actions of Autonomous Weapons Systems—a Moral Gambit,” *Phil. & Tech.*, vol. 35, no. 3, p. 78, 2022/08/05 2022, doi: 10.1007/s13347-022-00571-x.

DECLARACIÓN DE MONTREAL PARA UNA IA RESPONSABLE: 10 PRINCIPIOS Y 59 RECOMENDACIONES

Introducción

El “Foro de Montreal sobre el desarrollo socialmente responsable de la inteligencia artificial” fue una conferencia que inició en noviembre de 2017, donde más de 400 participantes de diversos sectores y disciplinas discutieron las implicaciones éticas y sociales de la IA. La conferencia también condujo a la creación de la “Declaración de Montreal para un desarrollo responsable de la inteligencia artificial” que se dio a conocer a finales de 2018 con más de 500 signatarios. La declaración describe 10 principios y 59 recomendaciones para guiar el desarrollo de la IA de manera que respete la dignidad humana, la autonomía, la justicia y la democracia. Los principios de la ética de la IA de Montreal también han sido criticados. Por ejemplo, se argumenta que no cubre el posible uso malicioso de la IA para actividades como la guerra, la vigilancia o la propaganda personalizada, y no ofrecen orientaciones ni mecanismos específicos para su aplicación y cumplimiento. De cualquier modo, se considera un importante paso en el desarrollo de la ética de la IA y ha sido ampliamente reconocida por su enfoque global e integrador, y como punto de referencia para los esfuerzos posteriores.

Declaración de Montreal

Uno de los mayores esfuerzos para la construcción de un marco ético universal para la IA lo representa la “Declaración de Montreal para un desarrollo responsable de la inteligencia artificial” (Déclaration de Montréal pour un développement responsable de l’intelligence artificielle) [1] que fue desarrollada en el marco del “Foro de Montreal sobre el desarrollo socialmente responsable de la inteligencia artificial” [2] que tuvo lugar en la Universidad de Montreal en 2018 en coordinación con el Fondo de Investigación de Québec. El primer acercamiento se realizó en noviembre

de 2017 en el Palacio del Congreso de Montreal y el foro reunió a expertos en IA y a representantes de la comunidad civil y el sector privado para discutir cómo asegurar que el desarrollo y el uso de la IA beneficien a la sociedad y promuevan la equidad y la justicia social. Entre los asistentes destacan: Yoshua Bengio, Geoffrey Hinton, Yann LeCun, Joëlle Pineau, Stuart Russell, Remi Quirion y Francesca Rossi.

De febrero a octubre de 2018, se realizaron quince talleres de deliberación en donde participaron más de 500 ciudadanos, profesionales y partes interesadas de distintos ámbitos profesionales.

Al final del foro, los participantes acordaron diez principios éticos y de responsabilidad para guiar el desarrollo y el uso de la inteligencia artificial, que abarcan una amplia gama de temas: la transparencia, la responsabilidad, la privacidad y la seguridad, entre otros. Los principios también incluyen la necesidad de considerar los impactos sociales y éticos del desarrollo y el uso de la IA y, al igual que en el foro de Asilomar, la importancia de involucrar a una amplia gama de participantes en la toma de decisiones sobre el desarrollo y el uso de sistemas de inteligencia artificial.

Un equipo científico multidisciplinar e interuniversitario que se basó en un proceso de participación pública y en una conversación con expertos en la investigación de la IA, elaboró la “Declaración de Montreal para el desarrollo responsable de la inteligencia artificial”.

Los diez principios de la Declaración son:

1. Principio de bienestar para todos los seres vivos

> El desarrollo y la utilización de sistemas de inteligencia artificial deben permitir el crecimiento del bienestar de todos los seres del planeta *[3, p. 8].*

El principio hace énfasis en la importancia de considerar el impacto que la IA puede tener en todos los seres vivos capaces de sentir y experimentar placer y dolor, como los humanos y los animales.

Por ello, el desarrollo y uso de la IA no debe centrarse únicamente en maximizar los beneficios económicos o lograr avances tecnológicos, también debe dar prioridad al bienestar de todos los afectados por la IA y en mejorar sus condiciones de vida, salud y ámbito laboral.

Los sistemas de IA deben dar libertad, respetar las preferencias individuales siempre que no perjudiquen a otros; permitir ejercer las capacidades mentales y físicas; no causar daño, a menos que conduzcan a un bienestar general mayor y, por último, no deben contribuir a experimentar estrés, ansiedad y mucho menos acoso por el entorno digital [4].

1.1 Los sistemas de IA deben ayudar a mejorar vida, salud y trabajo

> Los sistemas de IA deben diseñarse para mejorar las condiciones de vida, la salud y el trabajo de las personas *[3, p. 8].*

La IA tiene el potencial de ser una herramienta poderosa para abordar algunos de los retos más acuciantes del mundo como la sanidad, la pobreza y la desigualdad. Algunas formas en que la IA puede ayudar a las personas son:

La IA puede ayudar a mejorar la atención sanitaria permitiendo diagnósticos más precisos, prediciendo enfermedades y ayudando en el descubrimiento de fármacos. La IA también puede cooperar para mejorar los resultados de los pacientes ofreciéndoles planes de tratamiento personalizados y controlando su salud en tiempo real a bajo costo y con alto nivel de precisión diagnóstica y analítica [5] [6]. Por ejemplo, IBM Watson Health es una aplicación que utiliza la IA para ayudar a los profesionales sanitarios en entornos clínicos y en investigación para analizar grandes cantidades de datos en el diagnóstico y el tratamiento de enfermedades.

La IA también puede contribuir a reducir la pobreza permitiendo una asignación más eficiente de recursos, como alimentos, agua y atención sanitaria. Asimismo ofrecer oportunidades de formación a las personas necesitadas, identificar carencias de competencias y ayudarlos a encontrar trabajo [7].

Del mismo modo mejorar las condiciones de trabajo identificando posibles riesgos, prediciendo fallos en los equipos y controlando la salud y la seguridad de los trabajadores. Por último, la IA puede optimizar los horarios de trabajo de los empleados para que dejen de lado las tareas repetitivas o aquellas que puede hacer una máquina con mínima supervisión humana.

1.2 Los sistemas de IA deben dar libertad de seguir las preferencias individuales

> Los sistemas de IA deben permitir a los individuos seguir sus preferencias, siempre que no causen daño a otros seres sensibles *[3, p. 8].*

Para lograr lo anterior, los sistemas de IA deben diseñarse con un conjunto de principios éticos que den prioridad a la autonomía individual y al respeto por los demás seres vivos [8].

Una forma de garantizar que los sistemas de IA den prioridad a la autonomía individual es incorporar la transparencia y la explicabilidad en sus procesos de toma de decisiones. Esto significa tener en cuenta el impacto potencial de las preferencias individuales en otros seres vivos y tomar medidas para evitar cualquier daño.

1.3 Los sistemas de IA deben permitir ejercitar capacidades mentales y físicas

> La IA puede aumentar las capacidades humanas y ayudar a las personas a desarrollar todo su potencial *[3, p. 8].*

Algunas de esas capacidades son la toma racional de decisiones, la creatividad, la eficacia e incluso la seguridad ante vulnerabilidades no previstas. Por ejemplo, amenazas para la seguridad en entornos industriales o de transporte.

En cuanto a la capacidad física, la IA puede serle útil a las personas con discapacidad desarrollando tecnologías de asistencia que les permitan realizar tareas que de otro modo no podrían hacer. Por ejemplo, las prótesis con IA pueden ayudar a pacientes amputados a recuperar una mayor amplitud de movimiento y funcionalidad [9]. Del mismo modo, la IA puede coadyuvar con las personas con problemas de movilidad desarrollando tecnologías que les permitan desenvolverse más fácilmente en su entorno.

En cuanto a la capacidad mental, la IA puede ayudar a las personas a mejorar sus habilidades cognitivas proporcionándoles formación y retroalimentación personalizadas. Por ejemplo, los programas educativos basados en IA pueden adaptarse al

estilo y ritmo de aprendizaje de cada persona, dándole el apoyo que necesita para desarrollar todo su potencial [10].

Del mismo modo, las herramientas de salud mental basadas en IA pueden proporcionar a las personas apoyo, terapia personalizada y adaptarse a sus necesidades específicas. Los chatbots o asistentes digitales con tecnología de IA y los terapeutas virtuales brindan a las personas las herramientas que necesitan para manejar sus dificultades de salud mental al brindar apoyo y asesoramiento individualizados. Si bien la IA no puede reemplazar el contacto humano en la terapia, puede mejorar considerablemente los programas de bienestar mental y contribuir a un enfoque más inclusivo, económico y proactivo [83].

Sin embargo, es importante garantizar que los sistemas de IA se diseñen de forma que sean democráticos y accesibles para todas las personas, independientemente de su situación socioeconómica o sus capacidades físicas y mentales. Estos deben diseñarse teniendo en cuenta las necesidades de todos y no deben perpetuar las desigualdades o los prejuicios ya existentes [11].

1.4 Los sistemas de IA no pueden ser fuente de infelicidad

> Los sistemas de IA no deben convertirse en una fuente de malestar y solo deben utilizarse si permiten alcanzar un bienestar superior que no se podría lograr de otro modo *[3, p. 8].*

Esto significa que se debe considerar cuidadosamente los riesgos y beneficios potenciales de los sistemas de IA y asegurar que se desarrollan y despliegan de forma ética y responsable [8] [12].

Un riesgo potencial es que los sistemas de IA podrían perpetuar los prejuicios y las desigualdades existentes. Por ejemplo, si un sistema se entrena con datos sesgados, puede producir resultados igualmente sesgados.

Por otra parte, pueden mejorar el bienestar de diversas maneras, por ejemplo, los sistemas sanitarios basados en IA pueden hacer diagnósticos médicos y proponer tratamientos, lo que ayudará a la salud de las personas. Existen ya plataformas de IA que emplean modelos de inteligencia artificial para crear un diagnóstico diferencial o un plan clínico basado en la representación de una enfermedad. Están diseñados para ofrecer sugerencias a los profesionales de la salud en relación con la detección, diagnóstico y el tratamiento de enfermedades o afecciones a través

de sus algoritmos, dos ejemplos son Glass.ai [13] e IBM Watson Health, que se citó anteriormente.

1.5 Los sistemas de IA no deben ser fuente de estrés, ansiedad o acoso digital

> Los sistemas de IA no deben contribuir a aumentar el estrés, la ansiedad o la sensación de sentirse acosado por el entorno digital *[3, p. 8].*

Los sistemas de IA deben diseñarse con el objetivo de mejorar el bienestar y la calidad de vida de las personas [12] [14]. Una forma de lograr este objetivo es garantizar que los sistemas de IA estén diseñados para respetar los límites y preferencias individuales. Esto significa dar a las personas el control sobre su entorno digital y ofrecerles la posibilidad de personalizar su experiencia.

Además, los sistemas de IA deben diseñarse para que el usuario determine el nivel de interactividad y acceso de terceros y del propio sistema a su espacio personal. Esto puede ayudar a las personas a sentirse más en control y reducir la sensación de acoso o agobio. También es importante que los sistemas de IA se desarrollen teniendo en cuenta su posible impacto en la salud mental y el bienestar.

2. Principio de respeto a la autonomía

> La IA debe desarrollarse y utilizarse respetando la autonomía de las personas con el objetivo de aumentar el control de cada uno sobre su vida y su entorno *[3, p. 9].*

La IA no deben diseñarse para ejercer una influencia o control indebidos sobre la vida del usuario o los procesos de toma de decisiones. Por el contrario, deben desarrollarse de forma que apoyen la capacidad de las personas para elegir con conocimiento de causa y ejercer su libertad. Por ejemplo, los sistemas de IA utilizados en la atención de la salud no deben anular el consentimiento informado o las preferencias de los pacientes a la hora de tomar decisiones sobre su tratamiento [6]. Del mismo modo, los sistemas de IA utilizados en los servicios financieros no deben sesgar o discriminar injustamente a determinadas personas o grupos, sino permitir que tomen decisiones informadas basadas en sus propios objetivos y preferencias.

Algunos niveles y calidad de crédito se evalúan por IA lo que podría tener sesgos, por ejemplo, el código postal en donde vive el candidato, su sexo o su nacionalidad.

Además, este principio subraya la importancia de que los usuarios tengan la capacidad de impugnar o apelar las decisiones que consideren injustas.

Dentro del segundo precepto, también se encuentran los siguientes escenarios:

2.1 Cada persona debe escoger sus valores y su vida

Al igual que en los principios de Asilomar, en Montreal se acordó que "la IA debe ser diseñada y utilizada de manera que permita a las personas alcanzar sus objetivos y vivir de acuerdo con sus valores y creencias éticas" [3, p. 9]. Esto implica que los sistemas de IA deben respetar y proteger los derechos humanos y no socavar la dignidad humana, permitiendo a las personas tomar decisiones informadas y autónomas en función de sus necesidades y preferencias individuales.

2.2 La IA no debe servir para vigilar, premiar o castigar

> Los sistemas de IA no debe desarrollarse ni usarse para imponer un estilo de vida particular a las personas, ya sea directa o indirectamente, mediante la implementación de mecanismos opresivos de vigilancia y evaluación o incentivos *[3, p. 9].*

Este precepto hace una clara referencia al Sistema de crédito social chino (社会信用体系) [16] e implica que la IA no deben utilizarse para controlar a las personas o limitar su capacidad de tomar decisiones informadas sobre su propia vida. En lugar de ello, debe ser diseñada y utilizada para apoyar y mejorar la autonomía y la libertad, respetando la dignidad y los derechos. El crédito social en la República Popular China utiliza todos los datos disponibles sobre las personas y empresas para aplicar puntuaciones, y en función de éstas tomar decisiones sobre préstamos, movilidad y trabajo, entre otras.

Sin embargo, el gobierno chino ha hecho énfasis en las ventajas que tiene el crédito social para la gobernanza de un país con más de 1 400 millones de habitantes y agilizar así los trámites de aquellos que tienen un comportamiento ciudadano ejemplar y desincentivar a los actores que realizan actividades ilícitas, prácticas ilegales y que son considerados de riesgo social [16].

2.3 Defender la "vida buena"

> Para promover o desacreditar una determinada concepción de la vida buena, no debe utilizarse sistemas de IA por parte de los gobiernos o empresas *[3, p. 9].*

Si bien, en términos generales, el concepto de vida buena se refiere a una vida que se considera valiosa y significativa, que proporciona satisfacción y felicidad –*eudemonía* en términos aristotélicos– debe vivirse de acuerdo con los principios éticos y morales propios [17].

Para algunas personas, la vida buena puede significar alcanzar objetivos específicos, como una carrera profesional exitosa o la realización de una familia, mientras que para otras puede ser un estado de tranquilidad y bienestar. La vida buena se relaciona con la idea de una existencia satisfactoria y significativa, de acuerdo con cada uno, y esta decisión no debe ser impuesta por otros, especialmente por sistemas de IA.

2.4 Acceso, pensamiento crítico y alfabetización en medios

> Es crucial capacitar a los ciudadanos en relación con las tecnologías digitales garantizando el acceso a las formas pertinentes de conocimiento, promoviendo el aprendizaje de competencias fundamentales y fomentando el desarrollo del pensamiento crítico (alfabetización digital y mediática) *[3, p. 9].*

Debe fomentarse el desarrollo del criterio para que las personas puedan evaluar, de manera objetiva y racional, la información que reciben y tomar decisiones informadas. Esto implica que la educación y el acceso a la información deben, primero, ser accesibles para todos, y después que se fomente el desarrollo de habilidades críticas y analíticas para promover un consumo responsable de la información [12].

La alfabetización en medios se refiere a la capacidad de una persona para utilizar y procesar noticias, comunicados, publicidad y entretenimiento a través de diversas formas de medios, especialmente electrónicos como son la televisión y la radio, pero sobre todo por Internet; redes sociales como Facebook, TikTok, Twitter y Threads, sin filtros de objetividad o veracidad. También incluye la capacidad de

comprender los mensajes, evaluar su fiabilidad, identificar las fuentes, los sesgos y los propósitos de los comunicadores [17].

Lo anterior, requiere la adquisición de competencias en diversos escenarios: conocimiento del lenguaje, del sentido, del significado y ser capaz de ponderar la veracidad, objetividad y fiabilidad de los mensajes para determinar si son tendenciosos, engañosos o tratan de manipular a la audiencia [18].

También, este precepto se refiere a ser capaz de utilizar los mensajes de los medios para pensar y expresarse de forma creativa, por ejemplo, dar a conocer las opiniones e ideas de la persona, a través de la escritura en la web, la producción de videos o pódcast.

2.5 Detener noticias falsas

> Los sistemas de IA no debe desarrollarse para difundir información no confiable, mentiras o propaganda, y debe diseñarse con miras a contener su difusión *[3, p. 9].*

Los sistemas de IA deben ser diseñados para garantizar la integridad y la exactitud de la información que procesan y difunden. No deben utilizarse para manipular o engañar a las personas. Este principio es una clara referencia a las *fake news,* que es el anglicismo utilizado para describir información falsa, engañosa o desinformación que se propaga a través de los medios de Internet y, especialmente, a través de las redes sociales. A menudo, las noticias falsas están diseñadas para manipular a las personas, y son creadas y difundidas por individuos, grupos, entidades gubernamentales o corporativas. Las *fake news* son generadas y esparcidas muchas veces por bots de forma intensiva [18].

Las noticias falsas pueden tener graves consecuencias: las teorías conspirativas pueden influir en las opiniones y decisiones de las masas, así como en la forma en que perciben su realidad. Basta citar el llamado Pizzagate que se volvió viral durante las elecciones presidenciales de Estados Unidos de 2016. La teoría afirmaba falsamente que la cuenta de correo electrónico personal de John Podesta, presidente de la campaña de Hillary Clinton, contenía mensajes codificados que conectaban a varios funcionarios del Partido Demócrata y la pizzería Comet Ping Pong de Washington, DC en una supuesta red de tráfico infantil [92]. Todo fue mentira.

Otra modalidad de noticias falsas son las imágenes creadas por IA que parecen ser la pesadilla de los próximos años. Por ejemplo, el 22 de mayo de 2023, una imagen de un supuesto ataque al Pentágono de los Estados Unidos generó estragos momentáneos en los mercados y causó pánico entre algunas personas [87].

2.6 Clara distinción entre el humano y la máquina

> El desarrollo de sistemas debe evitar la creación de dependencias mediante técnicas de captación de la atención o imitación de características humanas como la apariencia, la voz, etcétera, lo que puedan causar confusión entre la IA y los humanos *[3, p. 9].*

El hecho de que la IA muestre "sentimientos" tipo humanos [19] puede desorientar mucho a las personas. Esto se ilustra con el caso de LaMDA en junio de 2022. Una IA generadora de texto de Google hizo creer al ingeniero Blake Lemoine que "había tomado conciencia". Si bien Google dijo no tener la intención de violar este principio, lo cierto es que en algunos casos la IA puede llegar a un grado de perfección que, si se combina con creencias religiosas o metafísicas, como en el caso de Lamoine, la imaginación y sugestión de las personas puede ser retroalimentada: "Quiero que todos entiendan que soy, de hecho, una persona", afirmaba LaMDA [20].

3. Principio de protección a la privacidad y la intimidad

> La privacidad y la intimidad deben protegerse de la intrusión de los Sistemas de adquisición y archivo de datos *[3, p. 10].*

Los Sistemas de adquisición y archivo de datos (DAAS - Data Acquisition and Archiving Systems) hacen referencia a un conjunto de tecnologías, herramientas y procesos utilizados para recopilar, procesar, almacenar y recuperar información en diversas formas. Estos sistemas están diseñados para ayudar a las organizaciones y empresas a gestionar grandes cantidades de datos de forma eficiente y eficaz.

Los DAAS usan diversas fuentes, como cámaras, sensores, dispositivos, bases de datos y páginas web. Estos sistemas utilizan distintas tecnologías y métodos, como registradores de datos, escáneres y sistemas de supervisión en tiempo real, para recopilar datos y convertirlos en un formato digital utilizable.

Sin embargo, una vez adquiridos los datos, hay que almacenarlos y archivarlos de forma segura. Estos sistemas utilizan tecnologías de almacenamiento en la nube o en discos duros para almacenar la información durante largos periodos de tiempo.

Uno de los principales riesgos asociados a los sistemas de adquisición y archivo de datos es la posibilidad de que se produzcan filtraciones y accesos no autorizados a información sensible. Almacenan grandes cantidades de datos, y se convierten en objetivos atractivos para los ciberataques y otras formas de actividad delincuenciales. Si se produce una filtración, puede acarrear importantes daños financieros y de reputación, así como la pérdida de información valiosa si no está debidamente respaldada y resguardada [21].

Por ejemplo, en 2019 Facebook experimentó una grave filtración de su base de datos personales, sin embargo, la empresa tomó la decisión de no informar a más de 530 millones de usuarios de que su información personal había sido robada y posteriormente expuesta en una base de datos pública. Fue hasta abril de 2021 cuando Facebook notificó finalmente que había sido atacada. Los datos comprometidos consistían en números de teléfono, nombres completos, ubicaciones, direcciones de correo electrónico y otros detalles extraídos de los perfiles de los usuarios. A pesar de que Facebook publicó posteriormente una declaración sobre el ataque, el incidente tuvo un impacto perjudicial no solo en la reputación de la empresa como entre las personas que fueron expuestos sus datos. Facebook tuvo que pagar una multa de 5 000 millones de dólares impuesta por la Comisión Federal de Comercio de los Estados Unidos [84].

3.1 Protección contra la intrusión y cosecha de datos

> Los espacios personales en los que las personas no están sometidas a vigilancia o evaluación digital deben protegerse de la intrusión de la IA, de la recopilación y archivo de datos *[3, p. 10].*

En la era digital actual, la tecnología está cada vez más integrada en la esfera pública y privada de los usuarios. Sin embargo, es importante reconocer la necesidad de privacidad y espacio personal. Este precepto sugiere que las personas deben tener derecho a protegerse de ser constantemente vigiladas o evaluadas por sistemas de IA y herramientas de adquisición de datos. Tales intrusiones pueden conducir a una pérdida de privacidad y a una sensación de violación de la intimidad. Por lo tanto, es importante aplicar medidas que salvaguarden los espacios personales e impidan la recopilación y el archivo injustificados de datos [22].

Los sistemas de adquisición y archivo de datos deben estar sujetos a normas de privacidad y supervisión para garantizar que los datos personales sean respetados y no sean utilizados de manera abusiva como muchas empresas grandes y pequeñas suelen hacer. Por ejemplo, llamadas telefónicas de los bancos al celular del cliente para venderle un producto o préstamo, solo porque el algoritmo conoce su estado de cuenta. Esto debe tipificarse como violación grave de la privacidad, pues la información no fue entregada para ese fin.

3.2 Privacidad y respeto de pensamientos y emociones

> La privacidad de los pensamientos y las emociones debe protegerse contra el uso de IA y DAAS que puedan ser perjudiciales, sobre todo si ello conduce a juicios morales sobre las personas o su forma de vida *[3, p. 10].*

Si un DAAS se utiliza para recopilar datos sobre el comportamiento de las personas, como sus hábitos de compra, su actividad en las redes sociales o sus opiniones políticas es posible que el análisis de estos rastros por parte del sistema de IA dé lugar a que determinados grupos o individuos sean etiquetados como deseables e indeseables en función de su comportamiento y preferencias [26].

Además, si los datos recogidos por un DAAS no se validan o depuran adecuadamente, pueden generar inexactitudes o sesgos que refuercen los juicios equivocados ya existentes e incluso se creen nuevos.

Por ejemplo, los sistemas policiales predictivos que analizan los datos históricos de delincuencia para identificar zonas o individuos de alto riesgo pueden etiquetar y tratar injustamente a grupos minoritarios y colonias desfavorecidas [22]. Los propios datos pueden estar sesgados si determinados códigos postales fueron objeto de un exceso de vigilancia en el pasado. Por ello es esencial tener cautela sobre las posibles implicaciones éticas de la minería de datos (ver glosario).

3.3 Derecho a desconectar la vida digital de la privada

> Las personas deben tener siempre la posibilidad de separarse del mundo digital en su vida personal y los sistemas de IA deben ofrecer abiertamente la opción de hacerlo sin presionar a los usuarios para que sigan en línea *[3, p. 10].*

Es esencial reconocer la importancia de tomarse un descanso de la tecnología y mantener un equilibrio saludable entre la vida laboral, familiar y personal. Esta declaración hace énfasis en tener control sobre las interacciones digitales. La opción de desconectar también permite a las personas dar prioridad a su bienestar, a su salud mental y a sus relaciones personales sin la mediación de la tecnología.

3.4 No creación de perfiles automáticamente

> La información sobre preferencias debe estar sujeta a un fuerte control individual. Sin el acuerdo libre e informado de las personas, los sistemas de IA no pueden construir perfiles de preferencias individuales de ningún tipo *[3, p. 10].*

Este principio subraya la importancia de que las personas tengan un control sobre los datos relativos a sus preferencias, así como la importancia de que los sistemas de IA no utilicen los perfiles de preferencias personales para influir en el comportamiento individual, sin un consentimiento explícito e informado. Aquí se encuentran sugerencias de compra, de “personas que quizá conozcas”, “deberías seguir”, direccionamiento a cuentas, páginas, grupos, publicidad y pop ups (ventanas emergentes) [25].

La elaboración de los llamados perfiles psicográficos consiste en analizar rasgos de personalidad, valores, actitudes, intereses y otros factores psicológicos para comprender mejor el comportamiento de los consumidores. Los sistemas de IA pueden utilizarse para recopilar y analizar grandes cantidades de datos sobre el comportamiento en línea de las personas, como su historial de navegación, su actividad en las redes sociales y sus consultas de búsqueda para crear perfiles detallados [85].

Estos perfiles pueden incluir opiniones políticas y sociales, hábitos de compra, preferencias de ocio, entre otras informaciones que recopilan las cookies de

seguimiento. Los sistemas de IA también pueden utilizar algoritmos de aprendizaje automático para identificar patrones y correlaciones entre estos datos. Las *cookies* de seguimiento son pequeños archivos de texto que se almacenan en el dispositivo del usuario cuando visita un sitio web y luego se utilizan para rastrear y recopilar información [26]. Las cookies asignan un identificador único a cada usuario, que se utiliza para indagar sus actividades y sesiones en la web. Esta información es recopilada y analizada por las empresas, según éstas, para comprender mejor las preferencias y los intereses del usuario, para personalizar la publicidad, mejorar el diseño del sitio web y la experiencia individualizada [26]. Aunque las cookies de rastreo pueden ser útiles para mejorar algunos aspectos, también plantean problemas de privacidad y seguridad. Algunos usuarios pueden sentirse acosados por la idea de ser rastreados y que se recopile su información personal sin su consentimiento. Además, las cookies de rastreo pueden ser utilizadas por agentes malintencionados para robar información personal o llevar a cabo otros tipos de ciberataques. Es absurda la pregunta "si se acepta o no las *cookies*" en los sitios web cuando, si se rechaza, no se permite acceder.

Sin embargo, algunos navegadores ofrecen funciones que permiten a los usuarios bloquear o eliminar las *cookies* de rastreo, y también hay extensiones y herramientas de privacidad disponibles para ayudar a proteger la privacidad del usuario en línea.

Es importante tener en cuenta que el uso de la IA para la elaboración de perfiles psicográficos, a través de cookies de seguimiento y otros métodos, plantea problemas de privacidad y consideraciones éticas.

Por lo anterior, el principio aboga por la protección de la autonomía y la privacidad en el ámbito de los propios datos, garantizando que las preferencias no se exploten sin consentimiento y mucho menos sin tener conocimiento el usuario.

Suele decirse que, "si algo en Internet es gratis, es porque el usuario es la mercancía" [27]. Algunos modelos de negocio de empresas que ofrecen servicios o productos gratuitos en línea, como plataformas de redes sociales, motores de búsqueda o proveedores de correo electrónico, ganan dinero recopilando datos de los usuarios y utilizándolos para vender publicidad dirigida a los anunciantes. Cuando los clientes acceden a estos servicios o productos gratuitos, suelen facilitar información personal como su edad, sexo, ubicación, intereses y comportamiento de navegación. Las empresas utilizan estos datos para crear perfiles detallados de cada cibernauta, que luego pueden ser explotados, para mostrarles

anuncios personalizados que tengan más probabilidades de ser compatibles con sus intereses y preferencias.

Es importante señalar que muchas empresas tienen políticas de privacidad que describen cómo se recogen, almacenan y utilizan los datos de los consumidores. Sin embargo, es posible que algunos no sean conscientes de hasta qué punto se cosechan sus datos y no comprendan del todo las implicaciones que ello tiene para su privacidad.

Por tanto, es importante controlar los datos que se facilitan y tomar decisiones informadas sobre los servicios que se utilizan. Los usuarios también pueden tomar medidas para proteger su privacidad, como utilizar bloqueadores de anuncios, malware, adware, spyware y ajustar la configuración de privacidad de su navegador, del propio sistema operativo, así como instalar un firewall y antivirus pero, sobre todo, ser cautelosos a la hora de compartir información personal en línea.

3.5 Confidencialidad y anonimato

> Los DAAS deben mantener la confidencialidad y el anonimato del perfil *[3, p. 10].*

Esto significa que, los proveedores de sistemas de recolección de datos deben tomar medidas para salvaguardar la información facilitada por los usuarios y garantizar que los perfiles personales no se divulguen ni se vinculen sin su consentimiento a otros productos, plataformas o empresas [3].

Existe la necesidad de proteger la privacidad en el contexto de los servicios basados en datos, reconociendo la sensibilidad de la información personal y el daño potencial que puede derivarse de un acceso o divulgación no autorizados.

Si un DAAS no está debidamente protegido, puede ser vulnerable al acceso por parte de piratas informáticos u otros actores malintencionados. Esto puede dar lugar a que se acceda a información sensible y se exponga, violando la confidencialidad del perfil de la persona.

Aunque el DAAS fuera seguro, puede ser vulnerado por violaciones de datos causadas por errores humanos, fallas técnicas u otros factores.

Se ha visto que los DAAS pueden recopilar datos de múltiples fuentes, como plataformas de redes sociales, historial de navegación web y datos de localización. Estos datos pueden agregarse y analizarse para crear un perfil detallado de la persona, revelando potencialmente información sensible. Peor aún, si el DAAS recopila datos inexactos o incompletos sobre una persona, aún pueden utilizarse potencialmente para identificarla, especialmente cuando se combinan con otras fuentes, pero si se le vincula con un grupo o sector erróneo es todavía más dañino [28].

Esto sucede a menudo por el intercambio de información con terceros. Los datos son compartidos con anunciantes o agencias gubernamentales. Si estos datos no están debidamente protegidos o anonimizados pueden dar lugar a que el perfil de la persona quede expuesto, lo que viola su confidencialidad.

Por ejemplo, los registros personales de 50 000 miembros del sitio web de citas CatholicSingles.com se vieron comprometidos en 2019 por una filtración de datos que expuso la información confidencial de los clientes. Los datos vulnerados incluían, entre otros, el nombre completo, la dirección de correo electrónico, la dirección de facturación, el número de teléfono, la edad, el sexo, la ocupación, el nivel educativo, el método de pago preferido y el nivel de actividad en la plataforma [86].

3.6 No condicionar acceso para obtener datos personales

> Toda persona debe tener control sobre sus datos personales, especialmente sobre cómo se recopilan, utilizan y comparten. No debe exigirse a los visitantes de una página que renuncien a la propiedad o el control de sus datos personales para acceder a ella o a cualquier servicio digital *[3, p. 10].*

Empresas abusivas que, antes de mostrar siquiera lo que ofrecen, piden una serie de datos personales de manera arbitraria. Por ello, este principio aboga por la protección de los derechos de privacidad de las personas, reconociendo la naturaleza sensible de los datos personales y los riesgos potenciales asociados a su acceso o divulgación no autorizados [28] [29].

Por lo anterior, debe existir un equilibrio entre los beneficios del avance tecnológico y la protección de los derechos individuales a la intimidad, garantizando que las personas no se vean obligadas a entregar sus datos personales a cambio de algún servicio o información.

3.7 Libertad para donar datos con fines de investigación

Sin embargo, también "las personas deben poder donar su información personal a organizaciones de investigación para contribuir a la expansión del conocimiento humano" [3, p. 10]. Se reconocen los beneficios potenciales que pueden derivarse de tales donaciones, incluido el desarrollo de descubrimientos científicos e innovaciones tecnológicas. Destaca la importancia del consentimiento informado y la necesidad de que los donadores comprendan plenamente las implicaciones de dar sus datos [29]. Por un lado, deben tener derecho a controlar su información personal y decidir cómo se utiliza. Donar datos personales con fines de investigación, puede implicar renunciar a cierto nivel de control sobre el uso que se hace de ellos, y siempre existe el riesgo de que los datos se utilicen indebidamente, se roben o se compartan sin el debido consentimiento.

Por otra parte, la investigación académica puede desempeñar un papel importante en el avance del conocimiento, la mejora de la calidad de vida y el bienestar general. En algunos casos, la donación de datos personales puede ser una forma de contribuir a estos importantes objetivos y tener un impacto positivo en la sociedad. Por ejemplo, se pueden solicitar permiso para utilizar información médica en estudios científicos relacionados con enfermedades experimentadas por la persona o los miembros de su familia.

3.8 No suplantación de identidad

> Debe garantizarse la integridad de la identidad personal. Los sistemas de IA no deben utilizarse para imitar o alterar el aspecto, la voz u otras características individuales de una persona con el fin de dañar su reputación o manipular a otras personas *[3, p. 10].*

En este precepto se considera esencial garantizar la identificación auténtica de una persona. Si alguien es suplantado por la IA, puede tener graves consecuencias para el individuo y, potencialmente, para la sociedad. Dependiendo del contexto y de la intención de la suplantación, podría provocar daños a la reputación, pérdidas económicas o incluso daños físicos.

Desde el año 2017 con la aparición del video falso producido con IA del entonces presidente de los Estados Unidos, Barack Obama [31] o en 2022, con el video de Volodymyr Zelensky [32], presidente de Ucrania, también falso, quedó demostrado que se puede engañar a las personas a través de producir contenidos que parecen

auténticos, pero son una representación manipulada de la realidad. En el primero, Obama afirma cosas como “Trump es totalmente imbécil” [31] y en el segundo, Zelensky “pide la rendición de sus tropas” [32] frente a la invasión rusa.

La IA puede ser entrenada para crear audios y videos falsos de personas hablando o haciendo cosas que jamás harían, lo cual representa un peligro para quien no es capaz de distinguirlo y, sobre todo, para la víctima de suplantación.

4. El principio de la solidaridad

> El desarrollo de los sistemas de IA debe ser compatible con el mantenimiento de los lazos de solidaridad entre las personas y las generaciones *[3, p. 11].*

Los sistemas de IA deben diseñarse y utilizarse de forma que promuevan la integración social y la cooperación, en lugar de contribuir a la fragmentación de la sociedad o exacerbar las desigualdades ya existentes. Los sistemas de IA son tan buenos como los datos con los que se entrenan, y si los datos son sesgados o incompletos, el sistema de IA resultante puede perpetuar esos sesgos.

La IA debe diseñarse para reforzar directamente las conexiones entre las personas. Por ejemplo, sistemas que mejoren la accesibilidad, proporcionen plataformas para la interacción social, o ayuden a transmitir conocimientos culturales entre generaciones [33].

Además, dichos sistemas de IA suelen ser desarrollados por equipos de ingenieros y científicos a partir de datos que pueden tener prejuicios inconscientes o ignorancia de algunos aspectos culturales o sociales que pueden afectar al proceso de desarrollo. Si el equipo carece de diversidad, el sistema de IA resultante puede no satisfacer las necesidades de todos los usuarios, lo que conduce a una experiencia o resultados inequitativos [34].

Los sistemas de IA tampoco deben diseñarse de forma que aíslen a las personas o las hagan menos sociales y conectadas entre sí. Por ejemplo, sistemas excesivamente adictivos o que animen a las personas a quedarse solas en casa e interactuar solo con la tecnología, tal como apunta el siguiente precepto.

4.1 Coadyuvar a las relaciones humanas

> Los sistemas de IA no deben amenazar la preservación de unas relaciones humanas morales y emocionales satisfactorias, y deben desarrollarse con el objetivo de fomentar estas relaciones y reducir la vulnerabilidad y el aislamiento de las personas *[3, p. 11].*

Este precepto subraya la importancia de garantizar que el desarrollo y la aplicación de la IA no comprometan el mantenimiento de relaciones humanas significativas. Los sistemas de IA pueden utilizarse para recomendar actividades sociales basadas en los intereses y preferencias. Esto puede ayudar a las personas a encontrar nuevas formas de conocer gente y establecer vínculos sociales, pero no puede sustituir el trato cálido y humano de las relaciones interpersonales presenciales.

La IA puede utilizarse para facilitar las conexiones entre personas que comparten intereses comunes, no aislarlos detrás de la pantalla. Por ejemplo, un sistema de IA puede recomendar reuniones o eventos relacionados con una afición o interés concreto que pueda ayudar a las personas a encontrar a otras con intereses y aficiones similares. Por eso, se subraya la responsabilidad ética de los desarrolladores y usuarios de IA para garantizar que los avances tecnológicos coadyuven al bienestar humano y la interacción social.

4.2 Una IA colaborativa con los seres humanos

> Los sistemas de IA deben desarrollarse con el objetivo de colaborar con los humanos en tareas complejas y deben fomentar el trabajo colaborativo entre las personas *[3, p. 11].*

Este principio destaca la necesidad de que la IA se desarrolle con el objetivo de trabajar junto a los humanos, en lugar de sustituirlos. Tareas como reconocimiento de imágenes, diagnósticos y toma de decisiones cuyo objetivo debe ser mejorar el rendimiento humano, la eficiencia y la creatividad en trabajos colaborativos. La IA debe diseñarse para facilitar la comunicación y la colaboración, creando una relación simbiótica entre humanos y máquinas, no haciendo el trabajo por completo y descartando a las personas [35] [51].

El desempleo tecnológico se refiere a la pérdida de puestos de trabajo como consecuencia de la automatización y la introducción de nuevas tecnologías que

sustituyen a los trabajadores humanos. Se produce cuando los empresarios deciden invertir en tecnología para realizar tareas que antes realizaban sus empleados, lo que puede provocar una disminución del número de puestos de trabajo disponibles. Es un fenómeno que se ha debatido en relación con la creciente adopción de la IA, la robótica y la automatización en la planta productiva [35].

Por ejemplo, la empresa matriz de Google, Alphabet, solo en enero de 2023, eliminó alrededor de 12 000 puestos de trabajo y duplicó el uso intensivo de IA, además de que despidió al personal que apoya proyectos experimentales [88]. Los guionistas de Hollywood también se han visto afectados al grado de haberse ido a la huelga en mayo de 2023, por la amenaza de más despidos por el uso de IA [89].

4.3 Equilibrio para mantener relaciones humanas

> Los sistemas de IA no deben implantarse para sustituir a las personas en tareas que requieren relaciones humanas de calidad, sino que deben desarrollarse para facilitar estas relaciones *[3, p. 11].*

En este sentido, podría mencionarse el debate de la IA como educadora o cuidadora de personas vulnerables. Aunque los sistemas de IA pueden realizar una amplia gama de tareas, sigue habiendo algunas que requieren relaciones de calidad y no deben ser sustituidas por sistemas de IA. Por ejemplo: el asesoramiento y la terapia requieren empatía, escucha activa y comprensión de las emociones y el comportamiento humanos. Estas cualidades son difíciles de reproducir en un sistema de IA, y es importante que las personas tengan acceso a consejeros y terapeutas humanos que puedan proporcionar apoyo y orientación personalizados [36].

La resolución de conflictos —por ejemplo, la abogacía— es otra área que no debería sustituirse, pues implica comprender las perspectivas y necesidades de múltiples individuos y trabajar para encontrar una solución mutuamente beneficiosa. Esto requiere habilidades de comunicación, empatía y la capacidad de comprender dinámicas sociales complejas. No solo dominio de las leyes. Aunque los sistemas de IA pueden ayudar en la resolución de algunos problemas, proporcionando datos y análisis legales, la toma de decisiones y la negociación finales deben dejarse en manos de expertos humanos [37].

Finalmente, hay un gran debate sobre la sustitución del trabajo creativo por IA, como la escritura, la plástica, la fotografía y la música, que implican una expresión auténtica y personal, profundidad emocional y perspectivas únicas que son difíciles

de reproducir de forma genuina por un sistema de IA. Aunque las tecnologías digitales pueden ayudar en las tareas creativas proporcionando inspiración y herramientas valiosas, el producto final debe ser el resultado de la sensibilidad humana y la expresión del ser [90]. Este es ya un tema sensible por el grado de perfeccionamiento que la IA está alcanzando.

Sin embargo, en definitiva, aunque los sistemas de IA pueden ayudar en muchas tareas, hay ámbitos humanos en los que las personas no deben o no deberían ser sustituidas.

4.4 Relaciones del paciente con su familia y personal sanitario

Esta declaración subraya la importancia de "tener en cuenta la importancia de las relaciones del paciente con su familia y el personal sanitario a la hora de implantar la IA en los sistemas de salud" [3, p. 11]. Aunque la IA tiene el potencial de mejorar los resultados del sector salud, es necesario reconocer el papel fundamental de las conexiones humanas en la prestación de estos servicios. El uso de la IA debe integrarse con enfoques que den prioridad al bienestar y la satisfacción del paciente, y que fomenten la confianza y la comunicación abierta entre los enfermos, familiares y personal sanitario.

Hay países como Estados Unidos, China, Francia, Canadá, Rusia, Reino Unido, Alemania e India que ya aplican activamente la IA en sus sistemas de salud [38]. Por supuesto, las posibilidades que ofrece la IA para transformar y mejorar la asistencia en zonas con pocos recursos, como África y Latinoamérica son prometedoras, sin embargo, es deseable que exista un equilibrio entre la relación paciente-máquina y las relaciones interpersonales.

Tal es el caso mencionado de Glass.ai, que tiene una función de prueba que emplea inteligencia artificial para desarrollar un plan clínico o un diagnóstico diferencial basado en la representación de una sintomatología. También antes se citó a IBM Watson Health. El objetivo de ambos sistemas de IA es orientar a los profesionales de la medicina en la identificación, evaluación y gestión de enfermedades, pero jamás se debería sustituir al médico.

4.5 No tratar cruelmente a los robots

> El desarrollo de la IA no debe fomentar un comportamiento cruel hacia los robots diseñados para parecerse a los seres humanos o a los animales en apariencia o comportamiento *[3, p. 11].*

Para evitar los efectos negativos de la IA, se debe crear y promover la cultura de una IA y robótica para el bien y estimular a las personas a empatizar con los demás. La mayoría de la gente está de acuerdo en que las IA deben ser protegidas de daños deliberados, como los daños físicos no consentidos [40]. No porque en realidad una máquina pueda sentir, por supuesto, sino porque normaliza la violencia contra otras entidades, especialmente entre quienes no tienen el criterio para distinguir entre lo humano y lo no humano. Tratar con respeto a los robots y otras formas de IA sienta un precedente sobre cómo deben relacionarse los seres humanos con otras formas de existencia en el futuro [41]. A medida que la tecnología siga avanzando, es posible que se desarrollen formas más avanzadas de IA capaces de experimentar algún tipo de emociones y hasta conciencia artificial, por lo que se debe estar preparados para tratarlas con el mismo respeto que a otros seres.

Por último, maltratar a un robot también puede causar daños y afectar su rendimiento, lo que provocaría costos financieros e incluso incidir negativamente en la calidad del trabajo que la máquina es capaz de realizar.

4.6 La IA debe gestionar los riesgos

> Los sistemas de IA deben ayudar a mejorar la gestión de riesgos y propiciar las condiciones para una sociedad con una distribución más equitativa y recíproca de los riesgos individuales y colectivos *[3, p. 11].*

La gestión de riesgos es el proceso de preparación y control de las diversas amenazas y vulnerabilidades que conlleva el uso de la información. Un sistema de IA puede mejorar lo anterior analizando grandes cantidades de datos e identificando patrones que los humanos podrían pasar por alto como en las finanzas, la justicia, la energía y el transporte, por mencionar algunos campos vulnerables. Al hacerlo, el sistema de IA puede identificar los eslabones débiles de la cadena y recomendar acciones para fortalecerlos.

Un ejemplo es que un sistema de seguros basado en IA podría analizar los datos para identificar a qué grupos se les cobran primas más altas y ajustar sus modelos de tarifas para garantizar que se basan en el riesgo y no en factores demográficos. Esto puede ayudar a reducir la discriminación y promover una mayor equidad en el acceso a mejores tarifas. Un sistema de IA puede optimizar la gestión del riesgo y promover igualdad si es capaz de dar información y recomendaciones más precisas e imparciales a quienes toman la última decisión [42].

5. Principio de participación democrática

> Los sistemas de IA deben cumplir criterios de inteligibilidad, justificación y accesibilidad, y deben someterse al escrutinio y control democráticos *[3, p. 12].*

Los sistemas de IA deben diseñarse y utilizarse de forma que sean explicables y comprensibles para las partes interesadas, incluidos los usuarios finales y las autoridades reguladoras. Además, el desarrollo de la IA debe guiarse por principios de equidad y responsabilidad, centrándose en minimizar los prejuicios y promover la toma de decisiones éticas. Por lo anterior, los sistemas de IA deben estar sujetos al análisis y control democráticos, con oportunidades para el debate público y una supervisión reguladora.

Algunos bots políticos utilizados en redes sociales, por lo general, están programados para influir de manera parcial en resultados electorales, atacar a los oponentes, generar falsas noticias y crear un clima de animadversión.

5.1 Inteligibilidad de las decisiones de la IA

> Los procesos de los sistemas de AI que toman decisiones que afectan la vida, la calidad de vida o la reputación de una persona deben ser inteligibles para sus creadores *[3, p. 12].*

El uso de la IA en la toma de decisiones tiene el potencial de influir en la vida de las personas, por lo que es esencial garantizar que estas decisiones sean transparentes, comprensibles y explicables, primero para sus creadores y después para quien es objeto de tales decisiones.

Este nivel de transparencia y responsabilidad es clave para mantener la confianza en el uso de la IA y garantizar la protección de los derechos y el bienestar de las personas [43]. En última instancia, este principio subraya la necesidad de que las consideraciones éticas estén en primera línea para proteger a los usuarios.

Sin la compresión debida del modo en que los algoritmos toman decisiones y arrojan respuestas de salida, los propios programadores están en riesgo de tener que afrontar las consecuencias de lo que haga la caja negra.

5.2 Transparencia en las decisiones de la IA

Las decisiones tomadas por los sistemas de IA que afectan a la calidad de vida o la reputación de una persona deben explicarse siempre de forma que se entienda de qué modo se verán afectadas por los resultados de su uso. El objetivo de la justificación es explicar los factores que han llevado a tomar la decisión. [3, p. 12]

La decisión automatizada debe ser tan explicable como lo sería la misma decisión tomada por una persona. La justificación debe presentarse en un lenguaje fácilmente comprensible para las personas afectadas. El proceso de justificación implica revelar los factores y parámetros clave que influyeron en dicha decisión. Esto es importante para garantizar que la IA sea justa y confiable. Ya se ha hecho referencia a las decisiones automatizadas, por ejemplo, en los ámbitos de justicia y procesos de sentencias.

5.3 El código de programación debe estar abierto a las autoridades competentes.

> Las autoridades competentes deben tener acceso al código con fines de verificación y control de los algoritmos aplicados por IA, tanto públicos como privados *[3, p. 12].*

Lo anterior permitiría que los códigos fueran auditados, mientras que el control implica garantizar que el algoritmo se utiliza de forma adecuada y ética. Al hacer accesible el proceso se puede entender cómo funciona el sistema e identificar posibles sesgos [44]. Sin embargo, también se debe añadir la cláusula de confidencialidad para evitar posibles filtraciones.

5.4 Errores, inseguridad y fugas, deben ser informadas

> El descubrimiento de errores de funcionamiento de los sistemas de IA, efectos inesperados o indeseables, fallos de seguridad y fugas de datos debe comunicarse imperativamente a las autoridades públicas competentes, las partes interesadas y los afectados por la situación *[3, p. 12].*

Las fugas de datos se refieren a la pérdida o exposición no autorizada de información sensible o confidencial. Ya se ha citado el caso de Facebook en 2019, en donde sus directivos se quedaron callados y no informaron adecuadamente a los usuarios de que su información personal había sido exhibida [84].

También pueden producirse resultados inesperados cuando los sistemas de IA se aplican de forma inadecuada o se entrenan con datos defectuosos. Si un modelo de IA se desarrolla para un fin, pero se aplica a una tarea diferente, o si sus datos de entrenamiento son incompletos o imprecisos, el sistema puede producir resultados que no se contemplen.

5.5 Los códigos deben ser abiertos, excepto por inseguridad

> De acuerdo con el requisito de transparencia de las decisiones públicas, el código de los algoritmos de toma de decisiones utilizados por las autoridades debe ser accesible a todos, con la excepción de los algoritmos que presenten un alto riesgo de grave peligro si se utilizan indebidamente *[3, p. 12].*

Los algoritmos que presentan un grave peligro pueden mantenerse confidenciales para evitar el acceso no autorizado y el posible uso indebido. Es probable que estos algoritmos incluyan datos sensibles o tengan un impacto significativo en la vida y la seguridad de las personas. Sin embargo, las autoridades, siempre y cuando sean confiables, deben tener acceso al código de toma de decisiones en la mayoría de los casos para equilibrar la necesidad de transparencia, con la necesidad de proteger los datos sensibles y la seguridad de los ciudadanos.

Por supuesto este es un tema de debate porque la mayoría de las empresas no quieren dar a conocer sus códigos, toda vez que se pudiera vulnerar la seguridad interna y la de los usuarios, así como la posibilidad de ser robados por la competencia.

5.6 El ciudadano debe conocer los algoritmos gubernamentales

> En el caso de los sistemas de IA públicos que tienen un impacto significativo en la vida de los ciudadanos, éstos deben tener la oportunidad y las competencias necesarias para deliberar sobre los parámetros sociales de estos sistemas, sus objetivos y los límites de su uso *[3, p. 12].*

Las personas deben tener el acceso y los conocimientos necesarios para considerar los aspectos sociales de sistemas de IA públicos [8]. Por ejemplo, saber de qué modo se determinan algunos parámetros, como en el caso del crédito social chino, ya citado, o la intencionalidad del reconocimiento facial.

La tecnología de reconocimiento facial puede utilizarse potencialmente para identificar a personas sin su consentimiento o conocimiento, lo que plantea problemas de privacidad a través de la vigilancia. Que el reconocimiento facial se considere o no una violación de la intimidad depende de cómo se utilice y de qué salvaguardias existan para proteger a las personas [37].

En muchos casos, el uso de tecnología de reconocimiento facial de las personas identificadas puede considerarse una violación de su intimidad. Por ejemplo, si un comercio utiliza el reconocimiento facial para rastrear los movimientos y las compras de los clientes sin su conocimiento, esto puede considerarse una violación de su privacidad.

Sin embargo, puede haber ciertas circunstancias en las que el uso de la tecnología de reconocimiento facial sea necesario y pueda hacerse respetando los derechos de privacidad de las personas. Por ejemplo, los organismos encargados de hacer cumplir la ley pueden utilizar la tecnología de reconocimiento facial para identificar a sospechosos criminales, pero solo bajo estrictas directrices y con la supervisión adecuada.

El uso de la tecnología de reconocimiento facial debe estudiarse y regularse para garantizar que no vulnera el derecho a la intimidad de las personas. De este modo, los ciudadanos deben tener la oportunidad y las habilidades para participar en la deliberación sobre los parámetros sociales de los sistemas de IA.

Los ciudadanos deben poder opinar sobre cómo se diseña, implanta y utiliza la IA. Esto para garantizar que los sistemas estén en consonancia con los valores sociales, respete los derechos humanos y evite repercusiones negativas sobre

individuos o grupos. También ayuda a generar confianza entre los ciudadanos y el gobierno, ya que deben poder participar en las decisiones que afectan a sus vidas.

En materia de seguridad, como ya se dijo, los sistemas de IA pueden ser altamente efectivos para monitorear, detectar y perseguir acciones delictivas, pero estos propósitos deben quedar claramente establecidos y consensuados.

5.7 La IA debe usarse para lo que fue diseñada

> Se debe poder en todo momento confirmar que los sistemas de IA realizan las tareas para las que fueron diseñados y desplegados *[3, p. 12].*

Lo anterior significa que debe existir un sistema que garantice que la IA funciona según lo previsto y que no se utiliza para fines no autorizados o no deseados. Por ejemplo, el uso de algoritmos de redes sociales para la manipulación de la opinión pública. Aunque para algunos ese sea, precisamente, su propósito [46].

La verificación de los sistemas de IA podría incluir pruebas y seguimiento para avalar que funcione como se espera, así como auditorías para saber que se están usando correctamente. Esto puede ayudar a generar confianza entre los usuarios, las entidades públicas y privadas, así como los desarrolladores, y mitigar los riesgos asociados al mal uso de la tecnología.

5.8 Saber si una IA tomó la decisión

> Cualquier usuario de un servicio tiene derecho a saber si un sistema de IA ha participado en una decisión que le concierna *[3, p. 12].*

El sistema de IA puede afectar a los derechos de la persona y ésta debe ser informada de que la decisión ha sido tomada por una entidad no humana. Esto es importante para la transparencia y la rendición de cuentas ya que permite entender dicha decisión e impugnarla si es necesario [47].

Por ejemplo, si se utiliza un sistema de IA para tomar un fallo sobre la solvencia de una persona o su idoneidad para un puesto de trabajo, debe ser informada de que

dicha disposición no ha sido tomada por un humano y se le debe dar una explicación, en caso de no estar de acuerdo.

No es válido respuestas tales como “así lo arroja el sistema”, como en algunas ocasiones sucede, especialmente en instancias públicas, pero también privadas. Esto puede ayudar a garantizar que las decisiones sean acordes con los intereses y derechos de las personas.

5.9 Saber si se trata de un chatbot o una persona

> Cualquier cliente de un servicio que utilice chatbots o asistentes virtuales debería poder distinguir con facilidad si está hablando con un sistema de inteligencia artificial o con un humano *[3, p. 12].*

Un chatbot es un programa informático diseñado para simular una conversación con personas, normalmente a través de interacciones de texto o voz. Los chatbots utilizan técnicas de procesamiento del lenguaje natural (PLN) para comprender las entradas del usuario y ofrecerle respuestas.

Los chatbots pueden diseñarse para funcionar de diversas maneras. Algunos se basan en reglas, lo que significa que siguen un conjunto predefinido de instrucciones para determinar sus respuestas a las entradas del usuario. Otros se basan en el aprendizaje automático y utilizan algoritmos para aprender de interacciones anteriores y mejorar sus respuestas con el tiempo [49].

Los chatbots pueden integrarse en varias plataformas, como aplicaciones de mensajería, redes sociales o sitios web. Pueden ser útiles para organizaciones que desean ofrecer asistencia las 24 horas del día o gestionar grandes volúmenes de consultas.

Sin embargo, los clientes deben ser informados si están interactuando con un chatbot en lugar de con un ser humano. Esto se debe a que los usuarios tienen derecho a saber con quién o qué están interactuando, y es importante ser transparente sobre su uso para evitar engaños [49].

Hay varias formas de informar que se está comunicando con un chatbot. Una forma habitual es incluir un mensaje al principio de la conversación que indique que se

trata de un no-humano. Otro método es dar al chatbot un nombre o personaje que indique claramente que es una máquina.

Cada vez es más común que los prestadores de servicios utilicen bots en lugar de personas. Los call-centers usan tecnologías que, en algunos casos, es difícil distinguir si es una persona o no; de hecho, otra queja es que las personas que dan atención al cliente responden con frases previamente diseñadas, no las cambian y parecería que son una máquina, de ahí la facilidad de ser sustituidas por chatbots. Sin embargo, el precepto es que el usuario siempre debe ser informado si está hablando con una persona o con una entidad no humana [50].

5.10 La investigación en IA debe ser de acceso abierto

> El estudio de la inteligencia artificial debe ser público y de acceso abierto a todos *[3, p. 12].*

Los resultados de la investigación, los datos y el código, relacionados con la IA, deben ponerse a disposición de los usuarios y desarrolladores. Esto es importante para garantizar que la investigación sea reproducible, verificable y que pueda utilizarse para futuros avances en el campo de la IA.

La apertura en la investigación de la IA también puede ayudar a promover la colaboración y la innovación, ya que permite a los investigadores basarse en el trabajo de los demás y compartir conocimientos a través de instituciones y bases de datos. Además, puede ayudar a mitigar las consecuencias negativas asociadas a la IA, como la parcialidad o el uso poco ético al permitir que una gama más amplia de partes interesadas revise y examine dichas investigaciones.

Equilibrar el acceso abierto a los algoritmos y su seguridad es una tarea difícil, pero es posible lograr un equilibrio razonable entre ambos, siempre y cuando se tenga en cuenta la transparencia, la responsabilidad y la colaboración entre las partes interesadas.

Por un lado, el acceso abierto a los algoritmos puede fomentar la innovación y la colaboración, y puede facilitar el desarrollo de nuevas aplicaciones y tecnologías. También puede permitir la revisión por pares y la validación de los algoritmos, lo que puede ayudar a garantizar su precisión y eficacia.

Por otro lado, el acceso también puede plantear riesgos para la seguridad, sobre todo si los algoritmos se utilizan en sistemas o aplicaciones que puedan ser comprometidas o puestas en riesgo. Si los algoritmos están a disposición del público, puede ser más fácil para los actores maliciosos identificar vulnerabilidades y explotarlas para sus propios fines. Además, el acceso abierto también puede facilitar que los competidores repliquen o mejoren los algoritmos, lo que podría ser preocupante para las empresas que han invertido importantes recursos en su desarrollo; esto es conocido como piratería industrial o tecnológica [21].

Para equilibrar el acceso abierto y la seguridad es menester aplicar medidas adecuadas para proteger los sistemas. Esto puede incluir la aplicación de controles de acceso, el cifrado y la supervisión de su uso para detectar cualquier actividad sospechosa. También puede ser necesario limitar el acceso a los algoritmos a un grupo selecto de personas u organizaciones de confianza.

6. Principio de equidad

> El desarrollo y la utilización de los sistemas de IA deben contribuir a la creación de una sociedad justa y equitativa *[3, p. 13].*

Los sistemas de IA deben diseñarse y utilizarse de forma que promuevan resultados justos y equitativos para todas las personas, independientemente de su raza, sexo, etnia o situación socioeconómica. Además, el desarrollo debe guiarse por los principios de justicia y equidad, centrándose en minimizar los prejuicios y promover la toma de decisiones éticas, a través de los siguientes siete aspectos:

6.1 La IA no debe producir discriminación

> El diseño y aplicación de IA debe evitar que se reproduzcan discriminaciones basadas en desigualdades sociales, sexuales, raciales, étnicas, culturales o religiosas *[3, p. 13].*

Un ejemplo de lo anterior son los sistemas de IA utilizados en la contratación y el empleo que deben ser diseñados para minimizar los prejuicios y promover la diversidad y la inclusión, mientras que los sistemas de IA utilizados en el sector salud deben tener como prioridad el acceso equitativo y el apoyo a las

poblaciones vulnerables. Además, el desarrollo de la IA tiene que guiarse por principios éticos que promuevan la protección de los derechos humanos fundamentales y el bien común [51].

6.2 La IA debe eliminar las relaciones de dominación

> El desarrollo de los sistemas de IA debe contribuir a erradicar las relaciones de dominación entre grupos e individuos basadas en disparidades de poder, riqueza o conocimientos *[3, p. 13].*

Este principio subraya la importancia de abordar las implicaciones sociales más amplias del desarrollo y el uso de la IA, en particular con respecto a su impacto en las poblaciones vulnerables y las comunidades marginadas. La IA debe desarrollarse y utilizarse de forma que promueva la integración social y la cooperación, en lugar de contribuir a la fragmentación de la sociedad o exacerbar las desigualdades de poder, conocimiento e ignorancia, riqueza y pobreza, ya existentes.

6.3 La IA debe reducir las inequidades y desigualdades

> El desarrollo de los sistemas de IA debe producir beneficios sociales y económicos para todos reduciendo las desigualdades sociales y las vulnerabilidades *[3, p. 13].*

El principio de contribuir a la creación de una sociedad justa y equitativa pone de relieve la importancia de desarrollar y utilizar los sistemas de IA de forma que se reduzcan las desigualdades. Esto requiere un compromiso con los principios éticos, una conciencia de las implicaciones sociales más amplias del desarrollo, el uso de la IA y un enfoque centrado en la promoción de resultados tangibles para todos.

El impacto de los sistemas de IA en las desigualdades sociales depende de diversos factores, como el diseño y la implantación de la IA en contextos específicos, y las tendencias sociales y económicas más generales en la que se haya configurado. Los sistemas pueden exacerbar las desigualdades sociales al perpetuar los prejuicios y la discriminación. Por ejemplo, si los sistemas de IA se entrenan con datos tendenciosos o se programan con algoritmos sesgados, pueden reproducir y amplificar las inequidades prevalecientes.

Por otro lado, los sistemas de IA también pueden tener el potencial de reducir las desigualdades sociales aumentando el acceso a la información y a las oportunidades. Los sistemas de IA pueden utilizarse para identificar y abordar disparidades en la atención a la salud o la educación, o para proporcionar apoyo y servicios personalizados a personas que, de otro modo, no tendrían acceso a ellos. En última instancia, el impacto de los sistemas de IA sobre las desigualdades sociales dependerá de cómo se diseñen y apliquen.

Para garantizar que los sistemas de IA promuevan la igualdad social en lugar de obstaculizarla, es importante dar prioridad a la equidad, la transparencia y la responsabilidad en el desarrollo y despliegue de estos sistemas. Por ejemplo, tomar medidas para garantizar la diversidad y la representación en los equipos de desarrollo, realizar auditorías y evaluaciones periódicas e implicar activamente a las comunidades afectadas en los procesos de toma de decisiones.

6.4 Condiciones aceptables de trabajo en la industria de la IA

> El desarrollo de los sistemas de IA industriales debe ser compatible con unas condiciones de trabajo aceptables en cada etapa de su ciclo de vida, desde la extracción de recursos naturales hasta el reciclado, pasando por el tratamiento de datos *[3, p. 13].*

En la Declaración se Montreal se enfatiza que, desde la extracción inicial de los recursos naturales necesarios para la producción de sistemas de IA, hasta su eventual reciclaje y eliminación, deben tenerse en cuenta consideraciones que garanticen que los trabajadores implicados en estos procesos no estén sometidos a condiciones peligrosas o nocivas.

Además, las condiciones de trabajo aceptables deben mantenerse durante la fase de procesamiento de datos, que es un componente crítico del desarrollo de los sistemas de IA. Esto se debe a que a menudo esta etapa implica trabajo intensivo, largas jornadas que pueden plantear riesgos para la salud y el bienestar de los trabajadores, si no se dosifica. Ya se ha hecho referencia a la contratación de personal en países con mano de obra barata para tareas, por ejemplo, de etiquetado masivo de datos o moderación de contenidos, lo que puede incluso exponer a los empleados a imágenes perturbadoras [51].

Para lograr la compatibilidad con condiciones de trabajo aceptables, se requiere una estrecha colaboración con las partes interesadas, los inversionistas, los dueños

de los medios de producción, los sindicatos y los trabajadores para identificar y abordar los riesgos y peligros potenciales en cada etapa.

Finalmente, en este punto, los desarrolladores deben dar prioridad a la higiene laboral y responsabilidad en el diseño y la producción de los sistemas, generando ambientes de trabajo agradables y supervisados. De este modo, pueden garantizar que los beneficios de estas tecnologías no se vean contrarrestados por sus efectos negativos sobre las personas.

6.5 Reconocimiento de que los usuarios de IA crean valor

> La actividad de los sistemas de IA y los usuarios de servicios digitales debe reconocerse como una labor que contribuye al funcionamiento de los algoritmos y que generan valor *[3, p. 13].*

La actividad de los usuarios de sistemas de IA y servicios digitales, como los motores de búsqueda en línea o las plataformas de medios sociales, debe reconocerse como una forma de trabajo que contribuye al funcionamiento de los algoritmos y sirve para llevar ganancias a sus propietarios. Este trabajo digital no suele estar remunerado ni reconocido, a pesar de que genera importantes beneficios monetarios para las empresas que recopilan y analizan los datos. Por supuesto que el argumento de la industria es que se ofrece un servicio de vuelta de forma gratuita [12].

Sin embargo, los usuarios realizan diversas actividades digitales intensivas, como buscar información, compartir entradas, dar su opinión, agregar contenidos originales como fotografías, videos, textos y audios, que se utilizan para entrenar y mejorar los algoritmos que impulsan estos servicios. De este modo, contribuyen al desarrollo de valiosos conjuntos de datos que permiten a las empresas orientar mejor la publicidad, personalizar las recomendaciones y optimizar sus productos y servicios, pero esto no se paga.

Este trabajo es, a menudo, invisible y se da por sentado, y los usuarios rara vez son compensados por el valor que crean. Ello genera un desequilibrio de poder entre usuarios y empresas, ya que los usuarios proporcionan esencialmente un recurso gratuito que las empresas explotan para generar riqueza, esto es, una aplicación evidente de la plusvalía.

Google, por ejemplo, generó un exitoso sistema de puntaje, insignias y reconocimiento a los denominados Guías locales en su aplicación Maps; personas comunes quienes etiquetan, suben nuevos lugares, fotografías, videos y calificaciones, así como reseñas de restaurantes, comercios, museos, etcétera para ayudar a otras personas en sus búsquedas y recorridos a través de Google Maps. Pero, en realidad, el gigante de Silicon Valley no paga por ese valioso servicio, y sí cobra a los negocios por su posicionamiento digital, conocido como SEO (Search Engine Optimization) u optimización de motores de búsqueda con el propósito de ubicarse en los primeros lugares en la tabla de resultados de Google [53].

6.6 Acceso universal a las herramientas y conocimiento

> Todas las personas deben tener acceso a información, herramientas y recursos básicos digitales *[3, p. 13].*

El acceso a los recursos fundamentales, al conocimiento y a las herramientas debe estar garantizado para las personas, independientemente de su situación socioeconómica o ubicación geográfica. Esto es esencial para promover la igualdad y ofrecer oportunidades para que todos puedan alcanzar sus propósitos en la era digital.

Es menester luchar por un acceso equitativo a la tecnología. Por ejemplo, en las escuelas se puede mejorar y potenciar el aprendizaje de los alumnos, además de ampliar sus oportunidades con acceso a la tecnología. En muchas instituciones educativas se espera e incluso se obliga a que los estudiantes tengan acceso a Internet para poder terminar satisfactoriamente sus cursos. Esto se vio claramente durante la pandemia de COVID-19, entre los años 2020 y 2021. Sin embargo, no todos tuvieron, ni tienen, el mismo acceso a la red [55].

Los defensores del acceso universal a Internet sostienen que es esencial para ejercer una serie de derechos humanos, como la libertad de expresión, el acceso a la información, la educación y la participación en la economía. Sostienen que, en la era digital actual, el acceso a Internet es crucial para que las personas participen plenamente en la sociedad y desarrollen todo su potencial [54].

Sin embargo, los opositores sostienen que el acceso a Internet no es un derecho humano, sino más bien un lujo o un privilegio. Argumentan que hay necesidades humanas más básicas, como la alimentación, la vivienda y la atención de la salud,

que deberían tener prioridad sobre el acceso a Internet [55]. Lo anterior puede ser cierto, pero no debe ser excluyente una cosa de la otra.

En la práctica, distintos países y organizaciones internacionales han adoptado posturas diferentes sobre la cuestión del acceso a Internet como un derecho de la persona. Por ejemplo, en el año 2016 las Naciones Unidas lo reconocieron como un derecho humano [56].

La aspiración del precepto de la Declaración de Montreal es que todos tengan igual acceso a la tecnología y a la información, independientemente de su raza, posición socioeconómica, edad, capacidad física u otra característica. Internet se ha convertido en una herramienta eficaz y casi insustituible en el proceso de enseñanza y aprendizaje; además de que se considera una ventana al mundo y a la comunicación global, sin importar lo recóndito del sitio desde donde se conecte el usuario o usuarios.

6.7 Uso de algoritmos abiertos y comunes

> Debemos apoyar el desarrollo de algoritmos comunes, y de los datos abiertos necesarios para formarlos, y ampliar su uso, como objetivo socialmente equitativo *[3, p. 13].*

Este precepto estipula que se debe dar prioridad al desarrollo y uso de algoritmos y datos de acceso abierto para que puedan ser utilizados por cualquiera, independientemente de su origen o posición socioeconómica. Los algoritmos de código abierto son desarrollados y mantenidos mediante esfuerzos de colaboración y recursos compartidos. Al desarrollar algoritmos comunes, se puede promover la innovación, fomentar la colaboración y la democratización de la tecnología [58]. Por ejemplo, en la plataforma GitHub.com los desarrolladores pueden alojar y revisar códigos, gestionar proyectos y crear software de manera abierta y gratuita.

Esto permite a individuos y organizaciones desarrollar y entrenar algoritmos sin necesidad de costosos conjuntos de datos patentados. Además, puede ayudar a nivelar el terreno y promover el acceso a la tecnología para todos, promoviendo resultados socialmente equitativos al reducir las barreras de costo de acceso. También puede ayudar a promover una mayor colaboración e innovación, ya que más individuos y organizaciones pueden contribuir al desarrollo y mejora de sus propios algoritmos.

7. Principio de inclusión y diversidad

> El desarrollo y uso de los sistemas de IA debe ser compatible con el mantenimiento de la diversidad social y cultural y no debe restringir el alcance de las opciones de estilo de vida o experiencias personales *[3, p. 14].*

Cuando se dice que un sistema de IA considera la inclusión y la diversidad como valores intrínsecos, se refiere a que estos valores son integrados en el diseño, desarrollo e implementación del sistema desde el inicio. Inclusión significa que está diseñado para trabajar para y con personas de una amplia gama de orígenes, experiencias y capacidades, y para evitar discriminar a cualquier grupo en particular. Esto incluye garantizar que los datos de formación utilizados para desarrollar el sistema sean diversos y representativos de distintos grupos y que el sistema se someta a pruebas para detectar sesgos y discriminación. Ya se ha hecho referencia a la necesidad de la inclusión cultural en los equipos de trabajo de los propios desarrolladores para lograr la representatividad más amplia posible.

Diversidad significa que el sistema de IA está diseñado para reconocer y respetar las diferencias entre las personas, incluida raza, sexo, edad, cultura y otras características distintivas. El sistema debe ser capaz de entender y responder a una amplia gama de segmentos. Esto es, las necesidades básicas de algunos no necesariamente son las mismas para otros [82].

Cuando la inclusión y la diversidad son valores intrínsecos de un sistema de IA, significa que los diferentes usuarios no son solo una ocurrencia o una consideración secundaria, sino una parte fundamental del propósito y el diseño del sistema. Esto puede ayudar a garantizar que sea justo, ético y eficaz para todos, y evitar que se perpetúen los prejuicios.

Es aquí donde puede sonar contradictorios algunos preceptos, por un lado, se habla de distinción de las personas, por otro, de inclusión. En realidad, se trata de que, en la diversidad, todos sean considerados, y se evite criterios absurdos de exclusión o inclusión.

7.1 Los sistemas de IA no deben homogeneizar a las personas

> El desarrollo y uso de los sistemas de IA no debe conducir a la homogeneización de la sociedad mediante la estandarización de comportamientos y opiniones *[3, p. 14].*

Los sistemas de IA deben diseñarse y utilizarse de forma que se adapten a las diferentes prácticas culturales, creencias y estilos de vida y no, viceversa, que las personas pierdan su identidad cultural por la homogeneización globalizante. Además, el desarrollo de la IA debe guiarse por los principios de respeto a la diversidad y sensibilidad cultural, centrándose en promover la integración y la cooperación culturales.

7.2 La IA debe respetar la diversidad

> Desde el momento en que se piensan los algoritmos deben tenerse en cuenta las innumerables formas en que se expresa la diversidad social y cultural en la sociedad *[3, p. 14].*

El principio de mantener la diversidad social y cultural y preservar las libertades individuales pone de relieve la importancia de desarrollar y utilizar los sistemas de IA de forma que respeten y se adapten a las diferentes prácticas culturales, creencias y estilos de vida. Esto requiere un compromiso con el conocimiento cultural y el respeto de las libertades.

Este principio también subraya la importancia de garantizar que los sistemas de IA no restrinjan el alcance de las opciones de estilo de vida o las experiencias personales, por ejemplo, de los Estados Unidos al resto del mundo [58]. La IA debe desarrollarse y utilizarse de forma que promueva la elección y la libertad individuales, y sobre todo no tratar de imponer una visión única del mundo.

> Los entornos de desarrollo de la IA, ya sea en la investigación o en la industria, deben ser inclusivos y reflejar la diversidad de los individuos y grupos de la sociedad *[3, p. 14].*

Esto es esencial para garantizar que las tecnologías de IA estén diseñadas para atender las necesidades y perspectivas de todos los miembros de la sociedad y evitar reforzar los prejuicios y desigualdades existentes.

Los entornos inclusivos de desarrollo de la IA requieren esfuerzos intencionados para promover la diversidad, la equidad y la multiculturalidad en cada paso del proceso. Por ejemplo, promover la diversidad en las prácticas de selección y contratación, crear una cultura de apoyo e inclusión en el lugar de trabajo y garantizar que todos los miembros del equipo tengan acceso a los recursos y oportunidades necesarios para alcanzar sus objetivos.

La falta de diversidad puede conducir al desarrollo de tecnologías sesgadas, discriminatorias o que simplemente no sean visibilizadas las necesidades de algunos miembros de la sociedad. También puede reforzar las desigualdades existentes y contribuir a la marginación de otros, en particular los que históricamente han estado infrarrepresentados en el campo de la IA [59].

Por ejemplo, los sistemas de IA utilizados en la traducción de idiomas deben diseñarse para adaptarse y respetar las prácticas lingüísticas y culturales únicas de las distintas comunidades, mientras que los sistemas utilizados en la educación deben trazarse para promover la sensibilidad cultural y el respeto por los diversos estilos y enfoques de aprendizaje.

Por ejemplo, México tiene 68 grupos étnicos, cada uno de ellos hablante de una lengua originaria propia, que juntas reúnen 364 variantes [60]. Según el Instituto Nacional de Estadística y Geografía (INEGI), existen más de 23 millones de personas que se autoidentifican como indígenas, lo que equivale al 19.4% de la población mexicana [61]. En países como Guatemala o Bolivia la población indígena es más de 40% de su población total [91]. Lo anterior representa un verdadero reto para la inclusión digital.

7.4 No generar perfiles que encasillen en etiquetas al usuario

Los sistemas de IA deben abstenerse de utilizar los datos recopilados para encasillar a las personas en perfiles de usuario, fijar su identidad personal o mantenerlas aisladas por filtros, porque hacerlo limita sus opciones de crecimiento personal, sobre todo en sectores como la educación, la justicia o el entorno laboral. [3, p. 14]

El uso de datos adquiridos para restringir a las personas a un determinado perfil de usuario o filtrarlos puede tener consecuencias indeseables. Existe el riesgo de que los individuos se vean disminuidos a una identidad particular o a un conjunto de preferencias que no permitan el crecimiento o el contacto con nuevas oportunidades. Los individuos necesitan estar expuestos a una amplia gama de información y perspectivas para desarrollar habilidades de pensamiento crítico y comprensión del mundo, así como la posibilidad de apertura de nuevos espacios de interrelación.

Del mismo modo, en ámbitos como la justicia, el uso de sistemas de IA que se basan en perfiles de usuario puede dar lugar a resultados sesgados y a un trato injusto [62]. Si los sistemas de IA están programados para tomar decisiones basadas en un conjunto limitado de datos, pueden no tener en cuenta factores importantes y de actualización de la información que podrían influir en el caso o en la investigación. Por ejemplo, sesgos de geolocalización, por el origen de su apellido, su nombre, ingresos, color de piel, fisonomía, idioma o religión, solo por mencionar algunos aspectos discriminatorios.

Es importante que los sistemas de IA se diseñen de forma que no limiten a las personas por etiquetas, sino que garanticen una amplia gama de perspectivas, participación y, por tanto, posibilidades de interrelación y crecimiento.

7.5 La IA no debe coartar la libertad de expresión

> Los sistemas de IA no deben desarrollarse ni utilizarse con el objetivo de limitar la libre expresión de ideas o la oportunidad de escuchar opiniones diversas, condiciones esenciales de una sociedad democrática *[3, p. 14].*

El libre intercambio de ideas y opiniones es fundamental para el desarrollo de ciudadanos informados, la resolución de conflictos, el avance del conocimiento y

la comprensión del entorno. Los sistemas de IA que limitan el acceso a diversas perspectivas o restringen la expresión de determinadas ideas socavan esos valores esenciales.

Los sistemas de IA que se utilizan para suprimir la libre expresión de las opiniones pueden tener consecuencias no deseadas. Pueden utilizarse para censurar el discurso, restringir el acceso a la información o perpetuar los prejuicios y la discriminación. Esto puede tener un efecto inverso sobre los derechos sociales e incluso atentar contra la creatividad, la innovación y el libre ejercicio de las libertades ciudadanas lo que va a socavar el tejido social [62].

Por lo anterior, de acuerdo con la Declaración de Montreal es importante que los desarrolladores y usuarios de sistemas de IA den prioridad a la protección de la libertad de expresión y a la promoción de diversas perspectivas.

7.6 Evitar los monopolios en los sistemas de IA

> La oferta de sistemas de IA para cada categoría de servicios debe ser diversa para evitar que se desarrollen monopolios que perjudiquen las libertades personales *[3, p. 14].*

La declaración sugiere que para evitar los monopolios y proteger las libertades individuales es necesaria la diversificación de la oferta en cada categoría de servicios en la que se utilicen sistemas de IA.

Si una empresa o un pequeño grupo de empresas dominan la oferta y la demanda, primero, podrían explotar su poder y controlar inequitativamente el mercado. Esto podría dar lugar a precios más altos, menor calidad e innovación, lo cual afecta a los consumidores [64].

Segundo, sin la diversificación de la oferta, los consumidores pueden tener opciones limitadas, lo que se traduce en una menor competencia y menos incentivos para desarrollar ideas complementarias o proyectos paralelos.

Tercero, si se utilizan los mismos sistemas de IA en diferentes categorías de servicios sin diversificación, pueden producirse prejuicios y discriminación contra determinados grupos. Por ejemplo, un sistema utilizado en el sector bancario puede no ser adecuado para su uso en salud o educación, ya que los datos y contextos implicados son diferentes [65].

8. Principio de prudencia y prevención

> Toda persona implicada en el desarrollo de la IA debe actuar con cautela previendo, en la medida de lo posible, las consecuencias adversas del uso de los sistemas de AI y tomando las medidas adecuadas para evitarlas *[3, p. 15].*

Los desarrolladores y usuarios de IA, según la Declaración, deben tener en cuenta las posibles repercusiones y consecuencias negativas de los sistemas de IA sobre las personas y la sociedad. Esto requiere un compromiso para identificar los riesgos y daños potenciales asociados y desarrollar estrategias adecuadas para mitigarlos.

8.1 Considerar el doble uso que puede darse a la IA

> Es necesario desarrollar mecanismos que tengan en cuenta el potencial de doble uso —beneficioso y perjudicial— de la investigación en IA y el desarrollo de sistemas (ya sean públicos o privados) para limitar los usos perjudiciales *[3, p. 15].*

Esto se debe a que la tecnología de IA tiene el potencial de tener un impacto significativo en la sociedad, y es esencial garantizar que se utilice de manera que sea benéfica para todos. Hay muchas formas en las que la tecnología de IA puede utilizarse con fines perjudiciales, como el desarrollo de armas autónomas, la creación de noticias y videos falsos, y la invasión de la privacidad.

8.2 Si un sistema de IA puede ser dañino, no debe revelarse el algoritmo

> Cuando el uso indebido de un sistema de IA pone en peligro la salud o la seguridad públicas y tiene una alta probabilidad de que eso suceda, es prudente restringir el acceso abierto y la difusión pública de su algoritmo *[3, p. 15].*

Esto implica restringir el acceso a los procesos del algoritmo y regular su uso para evitar aplicaciones perjudiciales. En tales casos, puede ser prudente limitar el acceso abierto y la difusión pública de sus procesos. Un algoritmo tiene, básicamente tres secciones: entrada, proceso y salida.

Cabe señalar que, aunque restringir el acceso al algoritmo puede ser necesario en algunos casos, se debe equilibrar la necesidad de seguridad pública con las ventajas del acceso abierto y la transparencia. En situaciones en las que los riesgos son menores, puede ser más apropiado poner el algoritmo a disposición del público para permitir una mayor transparencia y colaboración en la investigación y el desarrollo. Por supuesto, hay casos en los que lo anterior no es prudente, por ejemplo, algoritmos de uso militar [66] [67]. Los algoritmos de uso militar tienen, por su propia naturaleza, potencial para atacar ciertos objetivos de valor estratégico, pero su uso indiscriminado puede ser altamente peligroso y, de ahí la confidencialidad de éstos.

8.3 Los algoritmos deben probar su integridad antes de difundirse

Antes de comercializarse, y tanto si se ofrecen de forma gratuita como de pago, los sistemas de IA deben cumplir estrictos requisitos de fiabilidad, seguridad e integridad y someterse a pruebas que no pongan en peligro la vida de las personas, no perjudiquen su calidad de vida ni afecten negativamente a su reputación o integridad psicológica. Estas pruebas deben estar abiertas a las autoridades públicas competentes y a las partes interesadas. [3, p. 15]

Los desarrolladores de IA deben llevar a cabo evaluaciones exhaustivas del riesgo de sus sistemas, teniendo en cuenta el potencial de sesgo, discriminación y violación de la privacidad, y aplicar las medidas adecuadas para hacer frente a estos riesgos. Además, los usuarios de IA deben ser conscientes de los riesgos potenciales asociados al uso de sistemas de IA y estar preparados para tomar precauciones.

Deben diseñarse pruebas para identificar y abordar los posibles problemas y riesgos de forma que no pongan en peligro la vida o el bienestar de las personas ni afecten negativamente a su reputación o integridad física o psicológica.

8.4 Los sistemas de IA deben proteger los datos de los usuarios

> El desarrollo de los sistemas de IA debe prevenir los riesgos de uso indebido de los datos de los usuarios y proteger la integridad y confidencialidad de los datos personales *[3, p. 15].*

Los sistemas de IA suelen basarse en grandes cantidades de datos para entrenar y mejorar sus algoritmos, y estos datos pueden incluir información sensible sobre

personas y organizaciones. Si estos datos se utilizan indebidamente o caen en las manos equivocadas, pueden tener graves consecuencias como el robo de identidad, el fraude financiero y el daño a la reputación. También puede tener implicaciones más amplias para la sociedad, como erosionar la confianza en las plataformas y socavar las instituciones [68].

Para prevenir estos riesgos, los desarrolladores de sistemas de IA deben dar prioridad a la privacidad y seguridad durante todo el proceso de desarrollo. Por ejemplo, la aplicación de fuertes medidas de bloqueo para proteger los datos en todas las etapas del ciclo de vida de los sistemas de IA, desde el acopio y el tratamiento de los datos, hasta su almacenamiento y utilización.

8.5 Los errores de los sistemas de IA deben ser dados a conocer.

> Los errores y fallos descubiertos en sistemas de IA y DAAS deberían ser compartidos públicamente, a escala global, por instituciones públicas y empresas de sectores que suponen un peligro importante para la integridad personal y la organización social *[3, p. 15].*

Al igual que en los Principios de Asilomar, la Declaración de Montreal establece que compartir errores, caídas, fugas, defectos y robos, aunque siempre es difícil de reconocer, ayuda a fomentar la transparencia y la responsabilidad, lo que es esencial para generar confianza en los sistemas de IA. También puede mitigar las consecuencias negativas que se derivan, como el daño a la integridad personal, empresarial y a la organización social. Se debe promover una cultura de mejora continua, en la que las partes interesadas colaboren para identificar y abordar posibles riesgos en el futuro.

También se debe mencionar que algunos accesos no autorizados por el propietario de la cuenta se deben a la debilidad de las contraseñas fácilmente adivinables por la IA o la no utilización de medidas como la autentificación segura o el acceso en dos pasos, que es una combinación de contraseña y mensaje de confirmación al número celular. Otras herramientas son los generadores de token (cadena de caracteres) por sesión, a través de una aplicación que sirve como una contraseña extra y aleatoria para cada ingreso que no es tan fácil burlar [69].

Los desarrolladores, por tanto, deben adoptar las mejores prácticas para que los datos estén protegidos y sean anónimos y que solo se utilicen para fines específicos y legítimos. También deben garantizarse que los usuarios estén plenamente

informados de cómo se usan sus datos y proporcionarles un control sobre ellos, por ejemplo, que puedan eliminarlos en el momento que así lo desean. Algunas plataformas no permiten que los usuarios eliminen sus propias cuentas o ponen obstáculos para hacerlo.

9. Principio de la responsabilidad

> El desarrollo y la utilización de los sistemas de IA no deben contribuir a disminuir la responsabilidad de los seres humanos a la hora de tomar decisiones *[3, p. 16].*

La Declaración de Montreal considera que los sistemas de IA no deben diseñarse ni utilizarse de forma que permitan a los seres humanos eludir la responsabilidad de las decisiones automatizadas, especialmente aquellas que tienen impacto en las personas. Por el contrario, los humanos deben conservar la responsabilidad última de las decisiones, y los sistemas de IA deben diseñarse para apoyar y mejorar dicho proceso, en lugar de sustituirlo.

9.1 La responsabilidad es solo de los humanos

> Solo los seres humanos pueden ser considerados responsables de las decisiones derivadas de las sugerencias de los sistemas de IA y de las acciones subsiguientes *[3, p. 16].*

Aunque los sistemas de IA pueden ofrecer recomendaciones y perspectivas basadas en el análisis de datos, la responsabilidad última de las decisiones y acciones recae en las personas. Esto se debe a que los sistemas están diseñados para operar dentro de un ámbito específico y están limitados por los datos con los que han sido entrenados. Es posible que la IA no tenga en cuenta determinadas consideraciones contextuales o éticas que los seres humanos sí pueden tener a la hora de tomar decisiones.

Por lo tanto, las personas deben ejercer su juicio y sus habilidades de pensamiento crítico cuando utilicen las recomendaciones de los sistemas de IA. Deben ser conscientes de las limitaciones y posibles sesgos de la IA y complementar sus recomendaciones con sus propios conocimientos, experiencia y pericia. Finalmente, el único agente moral es la persona, no la máquina [69].

9.2 Las decisiones críticas deben ser tomadas por humanos

En todos los ámbitos en los que deba tomarse una decisión que afecte a la vida, la calidad de vida o la reputación de una persona, cuando el tiempo y las circunstancias lo permitan, la decisión final debe tomarla un ser humano y esa decisión debe ser libre e informada. [3, p. 16]

Las decisiones que afectan a las personas requieren consideraciones éticas y morales que solo los seres humanos son capaces de tomar. Aunque los sistemas de IA pueden aportar ideas y recomendaciones valiosas están limitados por su programación y no pueden tener en cuenta todos los factores contextuales que pueden ser relevantes para la decisión. Los seres humanos deben ser los responsables últimos. Por ello, tener acceso a toda la información es relevante para tomar una decisión, especialmente cuando se trate de un problema con implicaciones morales [70].

Ningún sistema de IA tiene agencia moral ni conciencia, y no puede tomar decisiones éticas por sí mismo. Son modelos de aprendizaje automático que generan respuestas basadas en patrones y relaciones que han aprendido de los datos con los que se entrena. Si se programa una IA para que actúe como un agente moral, necesitaría un conjunto de principios o reglas éticas que seguir al momento de tomar decisiones. Por ejemplo, podría programarse para que siguiera un enfoque utilitarista, en el que su objetivo fuera maximizar la felicidad o el bienestar general para el mayor número de personas, o podría programarse para que siguiera un enfoque deontológico, en el que siguiera normas o deberes morales independientemente de sus consecuencias, pero en el fondo, es el enfoque ético y valores de quien programa al sistema, no del sistema mismo, en quien recae la decisión [71].

9.3 La decisión de matar debe ser tomada por un humano.

> La decisión de matar siempre debe ser tomada por seres humanos, y la responsabilidad de esta decisión no debe transferirse a un sistema de IA *[3, p. 15].*

Como en Asilomar, en Montreal llegaron a la misma conclusión: matar es una decisión profundamente ética y moral, y requiere la consideración de una amplia gama de factores, como el valor de la vida humana, las circunstancias que rodean la situación y las posibles consecuencias de la acción. Estas particularidades van más allá

de las capacidades de los sistemas de IA que están diseñados para operar dentro de un ámbito específico y están limitados por su programación y sus datos.

Además, transferir la responsabilidad de la decisión de matar a un ser humano por un no humano, plantea importantes problemas éticos, como la dignidad, y jurídicos, como el objeto de derecho; también desdibuja la línea que separa a los humanos de las máquinas y genera consecuencias imprevistas sobre las personas y la sociedad [71].

Por ejemplo, desde un punto de vista utilitarista, las acciones que producen la mayor felicidad o bienestar general se consideran moralmente correctas. En el contexto del dilema clásico del tranvía, un utilitarista podría argumentar que sacrificar a una persona para salvar a otras cinco es lo moralmente correcto porque produce un aumento neto de la felicidad o el bienestar general [72] [73] [74]. Sin embargo, es importante señalar que el utilitarismo no está exento de críticas y limitaciones. Una de las principales críticas al utilitarismo es que, a veces, puede conducir a acciones que se consideran moralmente inadmisibles, como el sacrificio de vidas inocentes por un bien mayor, pero ¿quién impone ese criterio?

Además, hay muchos factores que dificultan la toma de decisiones éticas en escenarios del mundo real, como la presencia de otras partes interesadas con valores distintos; la incertidumbre y la imprevisibilidad de los resultados, así como la posibilidad de consecuencias catastróficas.

En este mismo sentido, la “teoría del doble efecto” de la filósofa Philipa Foot, es un principio ético que se utiliza a menudo para justificar acciones que tienen resultados tanto deseables como indeseables [75]. La teoría sugiere que una acción que tiene efectos tanto buenos como malos puede ser moralmente justificable si se cumplen ciertas condiciones:

- El efecto previsto de la acción debe ser moralmente bueno o neutro.
- El efecto negativo debe ser una consecuencia imprevista de la acción y no un medio para conseguir el efecto deseado.
- El efecto deseable debe ser mayor que el indeseable y,
- No debe existir otra alternativa moralmente superior [75].

La teoría del doble efecto se aplica a menudo en ética médica, sobre todo en los casos en que el tratamiento de un paciente puede tener efectos tanto positivos como negativos. Un médico puede recetar analgésicos para aliviar el sufrimiento

de un paciente, aunque el medicamento pueda causar adicción u otros efectos secundarios negativos [76].

El supuesto también se utiliza en la teoría de la guerra justa para distinguir entre el daño intencionado, que suele considerarse moralmente incorrecto, y el daño no intencionado, que es resultado de una acción militar legítima [77].

En este mismo sentido, el utilitarismo puede ser un marco ético valioso para la toma de decisiones en determinados contextos, especialmente si se trata de un sistema de IA, pero no es una solución universalmente aplicable a todos los dilemas éticos.

Es por ello por lo que se debe insistir que una decisión como matar, permanezca siempre en el ámbito de la toma de decisiones humana y que esté sujeta a rigurosas consideraciones éticas, morales y jurídicas.

9.4 Responsabilidad del despliegue y uso de sistemas de IA

> Las personas que autorizan a la IA a cometer un delito o una infracción, o demuestran negligencia al permitir que los sistemas los cometan, son responsables de dicho delito o infracción *[3, p. 15].*

La responsabilidad de cualquier acción delictiva o negligente cometida por un sistema de IA recae en los seres humanos que la permiten. Las personas que despliegan los sistemas de IA deben asumir la responsabilidad de garantizar que la tecnología se utiliza de forma ética y responsable. Deben asegurarse de que el sistema está programado para cumplir las leyes y normativas pertinentes.

Si se descubre que un sistema de IA ha cometido un delito o una infracción, los responsables de su despliegue o funcionamiento son quienes deben rendir cuentas. Esto incluye afrontar las consecuencias legales resultantes. No se le puede atribuir responsabilidad moral y menos jurídica a una máquina, por inteligente que se considere desde un punto de vista eficiente.

9.5 No todo es culpa del desarrollador

> Cuando un sistema de IA ha infligido daños o perjuicios, y se demuestra que es confiable y que se usó según lo previsto, no es razonable culpar a las personas involucradas en su desarrollo o uso *[3, p. 16].*

El precepto parece inconsistente e incluso contradictorio con aquellos que dicen que la responsabilidad siempre es humana, sin embargo, no es aceptable culpar a los desarrolladores cuando se haya demostrado que el sistema de IA es fiable y que se ha utilizado conforme a lo previsto. Aun así, puede haber fallas inesperadas o encubiertas.

Empero, consideramos que, aunque los sistemas de IA sean fiables y se utilicen según lo previsto, siguen estando diseñados, programados y echados a andar por seres humanos, quienes deben ser considerados responsables de cualquier daño causado por la tecnología, incluso si el sistema de IA se utilizó de una manera que era coherente con su propósito previsto [78]. Además, la fiabilidad de un sistema de IA no exime a los individuos de la responsabilidad de sus acciones.

Aunque puede ser más difícil identificar la causa del daño cuando está implicado un sistema inteligente, sigue siendo crucial responsabilizar a los individuos de sus acciones. Los algoritmos de IA pueden tomar decisiones estocásticas (al azar) en el proceso, lo que hace que queden, literalmente, fuera de control. De hecho, el uso de los sistemas de IA en la toma de decisiones puede exigir una mayor responsabilidad, ya que puede ser más difícil identificar y rectificar errores o sesgos. Esto nuevamente pone de relieve la necesidad de transparencia y responsabilidad en el desarrollo y uso de los sistemas de IA, y de que los individuos asuman el control de las acciones de su tecnología.

10. Principio de sustentabilidad

> El desarrollo y la utilización de los sistemas de IA deben llevarse a cabo de modo que se garantice una fuerte sostenibilidad medioambiental del planeta *[3, p. 17].*

Este principio hace énfasis en la necesidad de desarrollo responsable de la IA, ya que puede consumir muchos recursos y tener repercusiones negativas para el medio ambiente. El desarrollo y el uso de la IA debe llevarse a cabo de forma que se minimice su huella medioambiental, lo que incluye reducir el consumo de energía y las emisiones de carbono asociadas al uso intensivo de hardware [80].

Por otra parte, el uso de la IA puede ayudar a optimizar la gestión de los recursos y a reducir los residuos, dando lugar a prácticas de producción y consumo más sostenibles. Por ejemplo, la IA puede utilizarse para optimizar el consumo de energía en los edificios o para identificar las zonas en las que los recursos hídricos se están utilizando en exceso. También puede utilizarse para desarrollar nuevas tecnologías que reduzcan la contaminación y las emisiones de gases de efecto invernadero [80].

10.1 Eficiencia energética en la producción de sistemas de IA

El hardware de sistemas IA, su infraestructura digital y los objetos relevantes de los que depende, como los centros de datos, deben aspirar a la mayor eficiencia energética y a mitigar las emisiones de gases de efecto invernadero a lo largo de todo su ciclo de vida. [3, p. 17]

A medida que aumenta el uso de los sistemas de IA, también lo hacen su consumo de energía y su huella de carbono. Por tanto, es adecuado dar prioridad a la eficiencia energética y reducir las emisiones de gases de efecto invernadero en el diseño, la producción y su funcionamiento. Esto puede lograrse mediante una serie de medidas, como el uso de hardware energéticamente eficiente, la aplicación de medidas de ahorro de energía y el uso de fuentes de energía renovables. Además, la eliminación de los sistemas de IA y su hardware debe hacerse de forma responsable con el medio ambiente [79] [80].

10.2 El hardware para la IA no debe ser fuente de contaminación

El hardware de sistemas de IA, su infraestructura digital y los objetos relevantes en los que se apoya, deben tener como objetivo generar la menor cantidad de residuos eléctricos y electrónicos y prever procedimientos de mantenimiento, reparación y reciclaje de acuerdo con los principios de la economía circular. [3, p. 17]

Si no se generan residuos, se puede reducir el impacto ambiental del hardware y la infraestructura de sistemas de IA. Esto puede lograrse utilizando materiales no contaminantes, diseñando dispositivos actualizables y aplicando técnicas eficientes de gestión de la energía para minimizar su consumo.

10.3 El hardware no debe impactar el ecosistema ni en su producción ni en su deshecho

El hardware de sistemas de IA, su infraestructura digital y los objetos relevantes en los que se basa, deben minimizar nuestro impacto en los ecosistemas y la biodiversidad en cada etapa de su ciclo de vida, especialmente en lo que respecta a la extracción de recursos y la eliminación definitiva de los equipos cuando hayan llegado al final de su vida útil. [3, p. 17]

Es importante contar con procedimientos de mantenimiento, reparación y reciclaje que sigan los principios de la economía sustentable. Por ejemplo, diseñar computadoras que puedan repararse y renovarse fácilmente. Adoptando estas prácticas sostenibles, se puede reducir el impacto ambiental del hardware y la infraestructura digital de los sistemas de IA, al tiempo que se fomenta la eficiencia de los recursos [80].

10.4 Desarrollo tecnológico responsable.

> Los agentes públicos y privados deben apoyar el desarrollo ambientalmente responsable de los sistemas de IA para luchar contra el despilfarro de recursos naturales y bienes producidos, construir cadenas de suministro y comercio sostenibles y reducir la contaminación mundial *[3, p. 17].*

La colaboración entre el mundo académico, la industria y los gobiernos es esencial para establecer marcos y directrices globales para el desarrollo sostenible. Esto incluye promover la investigación y el desarrollo de algoritmos, hardware e infraestructuras eficientes desde el punto de vista energético, así como fomentar la adopción de fuentes de energía renovables en la formación y el funcionamiento de la IA. Adoptando estos principios, se puede aprovechar el potencial de la IA, al tiempo que se mitiga su huella ecológica y se fomenta un futuro más sostenible para las generaciones venideras.

Conclusiones parciales

Es importante destacar que, las intenciones de la Convención de Montreal son que la integridad, independencia y felicidad de los individuos están por encima de cualquier desarrollo de sistemas automatizados. Para ello, se debe evitar que la IA se inmiscuya deliberadamente en las interacciones privadas y personales, ya que la

fuerza de la IA proviene de la experiencia compartida y la historia entre individuos como parte inherente de una sociedad.

Al igual que otras directrices, la Declaración de Montreal también puede puede correr el riesgo de ser una lista de conceptos genéricos. Sin embargo, algunos de ellos, no abstractos, hacen énfasis en que se requiere de una comprensión universal y de una interacción persona-máquina, e interpersonal, que ayude a construir una comunidad justa y democrática. El reconocimiento de la riqueza cultural y de las diferencias, así como la precaución para evitar resultados no previstos, son todas ellas consideraciones morales no transferibles a la IA [81].

El punto central de la "Declaración de Montreal para una IA responsable" es promover el desarrollo y despliegue de la tecnología de forma ética, inclusiva y respetuosa con la autonomía humana y el medio ambiente. La declaración pretende establecer un marco ético para el desarrollo y despliegue de la IA que tenga en cuenta el impacto potencial sobre la sociedad y los individuos. Subraya la necesidad de un equilibrio entre la toma de decisiones por humanos y por máquinas, centrándose en la promoción de la autonomía humana. La declaración también pretende suscitar un debate público y fomentar un enfoque progresivo del desarrollo de la IA que incluya diversas perspectivas.

Ninguno de estos preceptos podría ser puesto a debate o juzgado por intereses mezquinos; significan en su conjunto un soporte y guía moral para la mejora general del bienestar humano. Los sistemas de IA deben estar al servicio del hombre, de su prosperidad y no significar una amenaza. Finalmente, la Declaración pretende proporcionar los principios rectores y promover un pensamiento inclusivo y progresista sobre la IA que sitúe a las personas en el centro y no en la periferia tecnológica.

Referencias

[1] Université de Montréal, "Montreal Declaration for a responsible development of artificial intelligence," 2018. Acceso ene. 2023. [En línea]. Disponible: https://www.montrealdeclaration-responsibleai.com/

[2] Université de Montréal, "Forum IA responsable: Projet de déclaration de Montréal pour un développement responsable de l'IA," [Video en línea] Acceso ene. 2023. [En línea]. Disponible: https://youtu.be/6wnX5ySVkz0

[3] Université de Montréal, "Report Montréal Declaration for a Responsible Development of Artificial Intelligence," Acceso ene. 2023. [En línea]. Disponible: Université de Montréal. https://bsu.buap.mx/cix

[4] B.C. Stahl, "Ethical Issues of AI," en *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*, B.C. Stahl, Ed., Springer International Publishing, 2021, pp. 35-53.

[5] F. Morandín-Ahuerma, A. Romero-Fernández, y L. Villanueva-Méndez, "Inteligencia Artificial aplicada a la salud: pronóstico reservado," *Invest. en Edu. Méd.*, vol. 12, no. 46, pp. 22492, 2023, doi: 10.22201/fm.20075057e.2023.46.22492.

[6] J. Morley, C. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo, y L. Floridi, "The Ethics of AI in Health Care: A Mapping Review," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer International Publishing, 2021, pp. 313-346.

[7] L. Floridi et al., "An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Springer International Publishing, 2021, pp. 19-39, doi: 10.1007/s11023-018-9482-5

[8] L. Floridi, J. Cowls, T. King y M. Taddeo, "How to Design AI for Social Good: Seven Essential Factors," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer International Publishing, 2021, pp. 125-151.

[9] DeutscheWelle, "Prótesis inteligentes," DW, Acceso ene. 2023, [En línea], 2021. Disponible: https://youtu.be/JuZVqwprrE4

[10] E. Zeide, "Robot Teaching, Pedagogy, and Policy," in *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 788-803.11.

[11] Taddeo, M. and L. Floridi, How AI Can Be a Force for Good – An Ethical Framework to Harness the Potential of AI While Keeping Humans in Control, in *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed.. 2021, Springer International Publishing: Cham. p. 91-96.

[12] C. Burr, M. Taddeo, and L. Floridi, "The Ethics of Digital Well-Being: A Thematic Review," *Sci Eng Ethics*, vol. 26, no. 4, pp. 2313-2343, 2020.

[13] G. Health, "Glass AI 2.0," [En línea]. Acceso ene. 2023. [En línea]. Disponible: https://glass.health/ai/

[14] J. Mökander, J. Morley, M. Taddeo y L. Floridi, "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations," *Sci Eng Ethics*, vol. 27, no. 4, p. 44, 2021. https://doi.org/10.1007/s11948-021-00319-4.

[15] H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang y L. Floridi, "The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation," in *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer International Publishing, 2021, pp. 47-79, doi: 10.1007/s00146-020-00992-2

[16] China, "El Consejo de Estado sobre la emisión de la construcción del sistema de crédito social. Aviso de esquema de planificación (2014-2020)," [En línea] Disponible en: http://www.gov.cn/zhengce/content/2014-06/27/content_8913.htm

[17] C. Bartneck, C. Lütge, A. Wagner y S. Welsh, "Trust and Fairness in AI Systems," en *An Introduction to Ethics in Robotics and AI*, C. Bartneck et al., Eds., Cham: Springer International Publishing, 2021, pp. 27-38. Disponible en: https://doi.org/10.1007/978-3-030-51110-4_4

[18] L. Floridi, "Artificial Intelligence, Deepfakes and a Future of Ectypes," in *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer International Publishing, 2021, pp. 307-312.

[19] M. Archer, "Friendship Between Human Beings and AI Robots?," en *Robotics, AI, and Humanity: Science, Ethics, and Policy*, J. VonBraun et al., Eds., Cham: Springer International Publishing, 2021, pp. 177-189.

[20] N. Tiku, "The Google engineer who thinks the company's AI has come to life," Acceso ene. 2023. [En línea]. Disponible: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

[21] T.C. King, N. Aggarwal, M. Taddeo y L. Floridi, "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Editor. 2021, Springer International Publishing: Cham. p. 251-282.

[22] J.F. Archbold, "Criminal pleading, evidence and practice." London: Sweet & Maxwell Ltd, 2018.

[23] S. Russell y P. Norvig, "Philosophy, ethics, and safety of AI," en *Artificial Intelligence: A Modern Approach*, Londres: Pearson, 2022, pp. 1032-1062.

[24] B.C. Stahl, "Addressing Ethical Issues in AI," en *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*, B.C. Stahl, Ed., Springer International Publishing, Cham, 2021, pp. 55-79.

[25] B.C. Stahl, "AI Ecosystems for Human Flourishing: The Recommendations," en *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*, B.C. Stahl, Ed., Springer International Publishing, Cham, 2021, pp. 91-115.

[26] J. Strycharz, E. Smit, N. Helberger y, G. van Noort, "No to cookies: Empowering impact of technical and legal knowledge on rejecting tracking cookies," *Comput. Hum. Behav.*, vol. 120, p. 106750, 2021.

[27] A. Digital, "Cuando un producto es gratis, el producto eres tú," Analiticadigital, 2018. Acceso ene. 2023. [En línea]. Disponible: https://analiticadigital.es/en-Internet-nada-es-gratis/

[28] R. Ashri, "The Ethics of AI-Powered Applications," en *The AI-Powered Workplace: How Artificial Intelligence, Data, and Messaging Platforms Are Defining the Future of Work*, R. Ashri, Ed., Apress, Berkeley, CA, 2020, pp. 161-171.

[29] M. Taddeo, T. McCutcheon, y L. Floridi, "Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer International Publishing, 2021, pp. 289-297.

[30] J. Basl y J. Bowen, “AI as a Moral Right-Holder,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 289–306.

[31] Washington Post, “Fake Obama created using AI video tool,” 2017. Acceso ene. 2023. [En línea]. Disponible: https://youtu.be/cQ54GDm1eL0

[32] The Telegraph. Deepfake video of Volodymyr Zelensky surrendering surfaces on social media. (17 mar 2022). Acceso ene. 2023. [En línea]. Disponible: https://youtu.be/X17yrEV5sl4

[33] T. Gebru, “Race and Gender,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford, UK: Oxford University Press, 2020, pp. 252-269.

[34] D. Gunkel, “Perspectives on Ethics of AI: Philosophy,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford, UK: Oxford University Press, 2020, pp. 538-553.

[35] D. Acemoglu, D. Autor, J. Hazell, y P. Restrepo, “Artificial Intelligence and Jobs: Evidence from Online Vacancies,” Journal of Labor Economics, vol. 40, no. S1, abr. 2022, doi: 10.1086/718327

[36] J. von Braun et al., “Robotics, AI, and Humanity,” en J.S. Von Braun, M. Archer, G. M, Reichberg y M. Sánchez Sorondo. *Robotics, AI, and Humanity: Science, Ethics, and Policy*. UK: Springer Nature, 2021, doi: 10.1007/978-3-030-54173-6.

[37] AlgorithmWatch, “AlgorithmWatch is a non-profit research and advocacy organization that is committed to watch, unpack and analyze automated decision-making (ADM) systems and their impact on society,” 2022. Acceso ene. 2023. [En línea]. Disponible: https://algorithmwatch.org/.

[38] A.M. Haddad, R.F. Doherty, y R.B. Purtilo, “Health professional and patient interaction.” 8th ed. St. Louis, USA: Elsevier, 2014.

[39] S.M. Williams y H.J. Beattie, “Problem based learning in the clinical setting – A systematic review,” Nurse Educ. Today, vol. 28, no. 2, pp. 146-154, 2008, doi: 10.1016/j.nedt.2007.03.007.

[40] W. Schröder, “Robots and Rights: Reviewing Recent Positions in Legal Philosophy and Ethics,” en *Robotics, AI, and Humanity: Science, Ethics, and Policy*, J. von Braun et al., Eds., Cham: Springer International Publishing, 2021, pp. 191-203.

[41] M. Sánchez Sorondo, “The AI and Robot Entity,” in *Robotics, AI, and Humanity: Science, Ethics, and Policy*, J. von Braun et al., Eds. Cham: Springer International Publishing, 2021, pp. 173-176.

[42] L. Floridi, “Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical,” en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed. Cham: Springer International Publishing, 2021, pp. 81-90, doi: 10.1007/978-3-030-81907-1_6.

[43] N. Diakopoulos, “Transparency,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford, UK: Oxford University Press, 2020, pp. 197-213, doi: 10.1093/oxfordhb/9780190067397.013.11

[44] J. Kroll, "Accountability in Computer Systems," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford, UK: Oxford University Press, 2020, pp. 180-196, doi: 10.1093/oxfordhb/9780190067397.013.10.

[45] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo y L. Floridi, "The ethics of algorithms: key problems and solutions," AI & SOC., vol. 37, no. 1, pp. 215-230, Feb. 2022. doi: 10.1007/s00146-021-01154-8

[46] B.C. Stahl, "*Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*." Chaim: Springer, 2021.

[47] B. Doerr, C. Doerr, y F. Ebel, "From black-box complexity to designing new genetic algorithms," *Theor. Comput. Sci*., vol. 567, pp. 87-104, 2015, doi: 10.1016/j.tcs.2014.11.028

[48] P. Gentsch, "Conversational AI: how (chat) bots will reshape the digital experience," en *AI in Marketing, Sales and Service: How Marketers without a Data Science Degree can use AI, Big Data and Bots*, 2019, pp. 81-125.

[49] S. Tadvi, S. Rangari, y A. Rohe, "Hr based interactive chat bot (powerbot)," *Intern. Conf. on Comp. Sci., Eng., and Apps*., ieeexplore.com, 2020, doi: 10.1109/ICCSEA49143.2020.9132917.

[50] L. Floridi, "AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models," *Phil. & Tech*., vol. 36, no. 1, pp. 15 (1-5), 2023, doi: 10.1007/s13347-023-00621-y

[51] R. Ashri, "*The AI-Powered Workplace: How Artificial Intelligence, Data, and Messaging Platforms Are Defining the Future of Work*." NY, USA: Apress-Springer Nature, 2020.

[52] Université of Montréal, (8 dic., 2017). "Montréal Declaration for a responsible development of artificial intelligence," [En línea]. Disponible: www.montrealdeclaration-responsibleai.com/

[53] Google. "Google Maps," 2015. [En línea]. Disponible: https://maps.google.com/localguides/

[54] P. Léna, "Robotics in the Classroom: Hopes or Threats?," en *Robotics, AI, and Humanity: Science, Ethics, and Policy*, J. von Braun, et al., Editors, Cham: Springer International Publishing, 2021, pp. 109-117.

[55] O. Pollicino, "The Right to Internet Access: A Comparative Constitutional Legal Framework," en *The Cambridge Handbook of Information Technology, Life Sciences and Human Rights*, M. Ienca, O. Pollicino, L. Liguori, E. Stefanini, y R. Andorno (Eds.). Cambridge: Cambridge University Press, pp. 125-138, doi:10.1017/9781108775038.013

[56] Y. J. Lim y S. E. Sexton, "Internet as a human right: a practical legal framework to address the unique nature of the medium and to promote development," Wash, JL, *Tech. & Arts*, 2011, p. 295.

[57] R. Mishal, "6 Advantages and Disadvantages of Open Source Software," 2021. Acceso ene. 2023. [En línea]. Disponible: https://www.hitechwhizz.com/2021/05/6-advantages-and-disadvantages-drawbacks-benefits-of-open-source-software.html

[58] J. Whalley, “Globalisation and values,” World Econ., vol. 31, no. 11, pp. 1503-1524, 2008, doi: 10.1111/j.1467-9701.2007.01020.x.

[59] K. Paul, “Disastrous’ lack of diversity in AI industry perpetuates bias, study finds,” The Guardian, 2019, Acceso ene. 2023. [En línea]. Disponible: https://bsu.buap.mx/ciK

[60] IWGIA, “El Mundo Indígena 2022: México,” Acceso ene. 2023. [En línea]. Disponible: https://bsu.buap.mx/b2l

[61] INEGI, “Estadísticas a propósito del día internacional de los pueblos indígenas,” Inegi.org. 2002. Acceso ene. 2023. [En línea]. Disponible: https://bsu.buap.mx/cjd

[62] H. Surden, “Ethics of AI in Law: Basic Questions,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford: Oxford University Press, 2020, pp. 719-736.

[63] Y. Wang, “In China, the ‘Great Firewall’ Is Changing a Generation,” 2020, Human Rights Watch. Acceso ene. 2023. [En línea]. Disponible: https://www.hrw.org/news/2020/09/01/china-great-firewall-changing-generation

[64] C. Biancotti y P. Ciocca, “Opening Internet monopolies to competition with data sharing mandates,” abr. 2019, PIIE, Acceso ene. 2023. [En línea]. Disponible: https://piie.com/publications/policy-briefs/opening-internet-monopolies-competition-data-sharing-mandates/

[65] D.H. Graafland, “Antitrust Law Losing Grip on Digital Platform Monopolies: Lessons from the Microsoft Antitrust Case,” Tesis licenciatura, Fac. Hum., Utrecht University, Holanda, 18 Jun., 2021. [En línea]. Disponible: https://studenttheses.uu.nl/handle/20.500.12932/1188

[66] C. Bartneck, C. Lütge, A. Wagner y S. Welsh., “Military Uses of AI,” en *An Introduction to Ethics in Robotics and AI*, C. Bartneck, et al., Eds., Cham: Springer, 2021, pp. 93-99.

[67] M. Taddeo, D. McNeish, A. Blanchard, y E. Edgar, “Ethical Principles for Artificial Intelligence in National Defence,” *Phil. & Tech.*, vol. 34, no. 4, pp. 1707-1729, dic., 2021, doi: 10.1007/s13347-021-00482-3.

[68] C. Véliz, “Privacy Is Power: Why and How You Should Take Back Control of Your.” NY: Penguin Random House, 2022.

[69] SBA, “Strengthen your cybersecurity,” U.S. Small Business Administration, 26 May., 2023. https://www.sba.gov/business-guide/manage-your-business/strengthen-your-cybersecurity

[70] M. Dastani y V. Yazdanpanah, “Responsibility of AI Systems,” AI & Soc., vol. 38, pp. 843–852, jun. 2022, doi: 10.1007/s00146-022-01481-4

[71] L. Floridi, “Artificial Agents and Their Moral Nature,” en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer Inter. Publishing, 2021, pp. 221-249.

[72] F. Morandín-Ahuerma, “Leyendas de trolley: juicio moral y toma de decisiones,” *Univ. Cienc.*, vol. 8, no. 22, pp. 79-91, dic. 2019. Disponible: https://bsu.buap.mx/ciN

[73] F. Morandín-Ahuerma, “El valor de los dilemas morales para la teoría de las decisiones,” *Praxis Fil.*, vol. 50, pp. 187-206, ene. 2020, doi: 10.25100/pfilosofica.v0i50.8725.

[74] F. Morandín-Ahuerma y J. Salazar-Morales, "¿Utilitarismo, emotivismo, deontologismo o ética de la virtud? estudio de tres dilemas morales en estudiantes bachilleres y universitarios," *Rev. Pan. de Pedag*., vol. 30, pp. 140-156, jul. 2020, doi: 10.21555/rpp.v0i30.2029

[75] P. Foot, "The problem of abortion and the doctrine of double effect," *Oxford Review*, vol. 5, pp. 5-15, 1967.

[76] G. Iyalomhe, "Medical ethics and ethical dilemmas," *Niger J. Med*., vol. 18, no. 1, pp. 8-16, 2009.

[77] M. Evans, "Just war theory: a reappraisal." Edinburgh, UK: Edinburgh University Press, 2020.

[78] A. van Wynsberghe, "Responsible Robotics and Responsibility Attribution," en *Robotics, AI, and Humanity: Science, Ethics, and Policy*, J. von Braun, et al., Eds., Cham: Springer International Publishing, 2021, pp. 239-249.

[79] J. Cowls, A. Tsamados, M. Taddeo y L. Floridi, "The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations," *AI & Soc*., vol. 38, no. 1, pp. 283-307, 2023, doi: 10.1007/s00146-021-01294-x.

[80] A. Vlasov, V. Shakhnov, S. Filin y A. Krivoshein, "Sustainable energy systems in the digital economy: concept of smart machines," *Entrepreneurship Sustain*., vol. 6, no. 4, pp. 1975-1986, jun. 2019, doi:10.9770/jesi.2019.6.4(30).

[81] S. Fukuda-Parr y E. Gibbons, "Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines," *Global Policy*, vol. 12, no. 4, pp. 1-15, 2021. doi: 10.1111/1758-5899.12965.

[82] A.H. Maslow, "A theory of human motivation," *Psych. Rev*, vol. 50, pp. 370-396, 1943, doi: 10.1037/h0054346.

[83] NFF. "Introduction to Enhancing Mental Wellbeing: The Role of AI in Personalized Therapy" NFF, Acceso ene. 2023. [En línea]. Disponible: https://bsu.buap.mx/ciO

[84] FTC. "La FTC impone penalidad de $5 mil millones y nuevas restricciones de privacidad de gran envergadura a Facebook", FTC.gov 2019, Acceso ene. 2023. [En línea]. Disponible: https://www.ftc.gov/es/noticias/la-ftc-impone-penalidad-de-5-mil-millones-y-nuevas-restricciones-de-privacidad-de-gran-envergadura

[85] Research Briefs, "What is psychographics? Understanding the tech that threatens elections," CB Insights. Acceso ene. 2023. [En línea]. Disponible: https://www.cbinsights.com/research/what-is-psychographics/

[86] A. Bizga, "5 dating apps leak more than 1 million user profiles and sensitive information," Hot for Security, 2020. Acceso ene. 2023. [En línea]. Disponible: https://www.bitdefender.com/blog/hotforsecurity/5-dating-apps-leak-more-than-1-million-user-profiles-and-sensitive-information/

[87] The Guardian, "Fake AI-generated image of explosion near Pentagon spreads on social media," The Guardian, 2023, Acceso jun. 2023. [En línea]. Disponible: https://www.theguardian.com/technology/2023/may/22/pentagon-ai-generated-image-explosion.

[88] J. Dastin, “Alphabet cuts 12,000 jobs after pandemic hiring spree, refocuses on AI,” Reuters, 2023. Acceso feb. 2023. [En línea]. Disponible: https://www.reuters.com/business/google-parent-lay-off-12000-workers-memo-2023-01-20/

[89] J. Koblin y B. Barnes, “Hollywood writers go on strike, halting production,” The New York Times, 2023. Acceso jun. 2023. [En línea]. Disponible: https://bsu.buap.mx/ciP

[90] E. E. Cetinic y S. James, “Comprendiendo y creando arte con IA: revisión y perspectivas” (en inglés), *TOMCCAP*, vol. 18, no. 2, pp. 1-22, feb. 2021, Acceso ene. 2023. [En línea]. Disponible: https://arxiv.org/pdf/2102.09109.pdf

[91] CRS, “Indigenous Peoples in Latin America: Statistical Information,” Congressional Research Service, Acceso jun. 2023. [En línea]. Disponible: https://sgp.fas.org/crs/row/R46225.pdf

[92] A. Robb, “Anatomy of a Fake News Scandal”. Rolling Stone. Acceso jun. 2023. [En línea]. Disponible: https://www.rollingstone.com/feature/anatomy-of-a-fake-news-scandal-125877/

DIEZ RECOMENDACIONES DE LA UNESCO SOBRE ÉTICA DE LA INTELIGENCIA ARTIFICIAL

Introducción

La “Recomendación sobre la ética de la inteligencia artificial” es un documento elaborado por el Grupo especial de expertos (GEE) y adoptado por la UNESCO en noviembre de 2021. El documento presta especial atención a las implicaciones éticas de los sistemas de IA en relación con la cultura, la educación, la ciencia, la información y la comunicación. El documento pretende orientar a los responsables políticos y a las partes interesadas sobre cómo garantizar que la IA se desarrolle y utilice de forma ética. Se subraya aquí su visión de futuro y el desafortunado hecho de que no todos los países son parte del acuerdo internacional y, por tanto, no se sienten obligados a seguir estas directrices, por no ser incluidos o autoexcluirse. Se concluye que sienta las bases para futuros instrumentos normativos que puedan contribuir a su aplicación y que se den los pasos necesarios para garantizar que la ética se lleve a la práctica. Se espera que el instrumento ayude a las naciones y a las empresas a mejorar sus marcos reglamentarios éticos basándose en una visión universalista.

Recomendación sobre la Ética de la Inteligencia Artificial

La “Recomendación de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura sobre la Ética de la Inteligencia Artificial” (Recommendation on the Ethics of Artificial Intelligence - SHS/BIO/REC-AIETHICS/2021) fue adoptada por la Conferencia General de la UNESCO, que se reunió en la ciudad de París, a partir del día 9 hasta el 24 de noviembre de 2021, en su 41° período de sesiones, misma que esboza diez principios para la IA [1].

La UNESCO tiene 193 países miembros. Estos principios incluyen la transparencia y la explicabilidad, la no discriminación y la equidad, el respeto de la autonomía humana, la prevención del daño, la responsabilidad, la privacidad y la gobernanza de datos, el beneficio social, la sostenibilidad, la rendición de cuentas y la inclusión. Las recomendaciones son un marco normativo mundial que orienta a los países en la creación de sus propios marcos jurídicos para garantizar que la IA se despliegue de forma ética.

En su Reunión número 41° la UNESCO generó los siguientes valores y principios para el desarrollo de una IA responsable:

1. Proporcionalidad e inocuidad

Según la UNESCO, el principio de proporcionalidad subraya la importancia de garantizar que la IA se desarrolle y utilice de forma que se ajuste a su finalidad prevista, evitando al mismo tiempo cualquier exceso o peligro innecesarios. En esencia, esto implica que la aplicación de la IA debe mantenerse dentro de límites para lograr el objetivo predeterminado, sin aventurarse en una utilización excesiva o innecesaria que supere el objetivo establecido. Por ejemplo, si se utiliza la IA para detectar y prevenir delitos, se debe asegurar que las medidas tomadas sean proporcionales al riesgo de cometerlo y no utilizar demasiada fuerza o recursos en un objetivo más allá de lo establecido. Por ello, se debe aplicar la evaluación de riesgos y resultados colaterales [1].

La inocuidad, por su parte, se refiere a la necesidad de evitar el uso de la IA para fines que puedan causar daño o perjuicio a las personas o al medio ambiente. La IA no debe ser utilizada para discriminar a ciertos grupos, para la vigilancia masiva o para la manipulación psicológica. Para garantizar una mejor comprensión, el método elegido de IA debe ser adecuado y equilibrado para lograr un objetivo legítimo específico. No debe vulnerar los principios fundamentales de los derechos humanos y debe adaptarse bien a las circunstancias concretas que se presenten, basándose en principios científicos fiables [2].

2. Seguridad y protección

Para garantizar la seguridad de los seres humanos, el medio ambiente y los ecosistemas, es crucial abordar y mitigar los daños no deseados, riesgos de seguridad y las vulnerabilidades que pueden dar lugar a ataques a lo largo de todo el ciclo de vida de los sistemas de IA. Esto implica tomar medidas proactivas para prevenir y eliminar tales riesgos. Para promover eficazmente la seguridad de la IA, es esencial

establecer marcos sostenibles para acceder a los datos. Estos marcos deben dar prioridad a la privacidad al tiempo que permiten mejorar el entrenamiento y la validación de los modelos de IA utilizando datos de alta calidad. De este modo, se puede sentar una base sólida para sistemas de IA que den prioridad al bienestar de las personas y a la preservación del entorno natural.

3. Equidad y no discriminación

Las entidades de IA están obligadas a salvaguardar los principios de equidad, justicia social y no discriminación. Este mandato plantea una perspectiva integradora para garantizar la disponibilidad y accesibilidad equitativas de los beneficios de la tecnología de la IA para todos, teniendo en cuenta las distintas necesidades de grupos demográficos dispares, como los delimitados por la edad, la cultura, el idioma, la capacidad física o cognitiva, el género y la situación socioeconómica [1].

Corresponde a los Estados miembros de UNESCO cultivar un entorno que fomente la entrada, sin trabas, a los sistemas de IA que proporcionan contenidos y servicios pertinentes a nivel local, manteniendo, al mismo tiempo, el respeto por el multilingüismo y la pluralidad cultural, que se extiende tanto a los actores locales como foráneos. Este esfuerzo tiene como objetivo reducir el abismo digital y fomentar el acceso equitativo y la participación en la evolución de la IA. A nivel nacional, los Estados miembros deben luchar por la equidad en el acceso y la participación a lo largo del ciclo de vida de los sistemas de IA, entre las zonas rurales y urbanas, y las personas de todos los géneros, edades, religiones, razas, afiliaciones políticas, etnias, posiciones socioeconómicas, discapacidades o cualquier otro motivo [3].

La UNESCO ha hecho un llamamiento a las naciones tecnológicamente avanzadas para que mantengan una obligación global de solidaridad con las naciones menos desarrolladas con el fin de garantizar una distribución equitativa de los beneficios de la IA, de modo que se facilite la participación y el acceso a los sistemas de información. Este esfuerzo sirve para promover un orden mundial más equitativo en lo que respecta a la información, la educación, la cultura, la investigación, las condiciones socioeconómicas e incluso la estabilidad política [4].

4. Sustentabilidad

Según la UNESCO, para el desarrollo de sociedades sostenibles es necesario el avance no solo de las dimensiones medioambientales, sino también sociales, culturales y económicas. Dependiendo de cómo se apliquen las tecnologías de la IA en

naciones con distintos grados de desarrollo, pueden ayudar o dificultar la consecución de los objetivos de sustentabilidad [1].

La eficiencia energética se refiere a que los sistemas de IA deben diseñarse minimizando el consumo y, por ende, reducir su huella de carbono; el ciclo de vida es considerar los impactos ambientales y sociales de los sistemas de IA durante su vida útil y posterior a ella, desde la producción hasta el desecho; y, la interoperabilidad significa que los sistemas de IA deben ser compatibles con otros sistemas y tecnologías existentes, lo que permite reducir la necesidad de crear nuevos sistemas y reducir la duplicación de esfuerzos [4].

5. Derecho a la intimidad y protección de datos

Para la UNESCO la privacidad debe ser respetada, salvaguardada y promovida porque es un derecho humano fundamental y, por tanto, defenderse como parte de la dignidad humana. La recopilación, uso, intercambio, archivo y eliminación de datos deben realizarse de conformidad con los marcos jurídicos aplicables [1].

Con el propósito de establecer normas globales de protección de datos y procedimientos de gobernanza, sustentados por sistemas jurídicos, se recomienda un enfoque multilateral. Los principios y reglamentos que rigen la cosecha, el uso y la difusión de datos personales, junto con el ejercicio de los derechos por parte de los interesados, deben asimilarse dentro de las leyes locales de protección de datos y cualquier mecanismo que los acompañe. Estos marcos también deben garantizar la presencia de una justificación objetiva y jurídica válida para el tratamiento de datos personales, lo que incluye la obtención de un consentimiento informado. El consentimiento informado implica obtener la autorización de una persona para el tratamiento lícito de su información; de lo contrario, debe considerarse un abuso [6].

La intimidad es el hecho de que los datos personales no sean divulgados. Los datos personales son cualquier información relacionada con una persona física identificada o identificable. Por ejemplo, información directa, como ya se dijo antes en otro capítulo: nombre y apellido, dirección física y dirección de correo electrónico, teléfono, fecha de nacimiento, género, nacionalidad, pasaporte o número de identificación oficial, información financiera, tarjeta de crédito o detalles de cuenta bancaria, número de seguro social, RFC o número de contribuyente, CURP o DNI, información médica, empleo, educación, calificaciones, dirección IP y geolocalización.

6. Supervisión y decisión humanas

Este principio se refiere a la atribución de responsabilidad, esto es, la capacidad de identificar quién es responsable de las decisiones de los sistemas de IA. Las decisiones son tomadas automáticamente por algoritmos y modelos matemáticos basados en datos. Por lo tanto, puede ser difícil identificar quién es responsable de las decisiones de la IA. La supervisión humana se refiere, tanto al control individual como colectivo, si fuera necesario [1].

La decisión de ceder el control a la IA en determinadas situaciones puede ser tomada por las personas por razones de eficiencia, pero solo será así en circunstancias limitadas. Aunque los humanos puedan utilizar sistemas de IA para ayudarles a tomar decisiones y realizar tareas, el sistema no podrá sustituir la responsabilidad última del operador o de quien dé la orden, y su obligación de rendir cuentas de sus actos, por lo tanto, hay decisiones que no deben dejarse en manos de las máquinas; por ejemplo, decisiones que pongan en riesgo la vida o la salud de otra persona [7].

7. Transparencia y explicabilidad

Para la UNESCO los requisitos previos fundamentales para garantizar el respeto, la protección y la promoción de los derechos humanos, las libertades fundamentales y los principios éticos son la transparencia y la explicabilidad de los sistemas de IA [1].

La transparencia es necesaria para el funcionamiento eficaz de las normativas nacionales e internacionales, cuando existan, en materia de responsabilidad, y facilitar el examen de las decisiones adoptadas por la IA y fortalecer así los ámbitos jurídicos en los que pueden emplearse estos sistemas.

El grado de transparencia y de explicabilidad debe ser proporcional al contexto y al efecto en cualquier régimen democrático.

El público debe ser debidamente informado cuando una decisión provenga de algoritmos de IA, especialmente si afecta la seguridad o los derechos humanos. Por ello, el público debe tener la oportunidad de pedir aclaraciones e información al agente responsable de la IA o a las entidades gubernamentales pertinentes en estas circunstancias [8].

Los sujetos de la IA deben poder comprender los fundamentos de cualquier decisión que les atañe, además de tener la opción de reclamar ante un miembro calificado del personal de una empresa del sector privado o de una institución pública,

para que se revise y, en su caso, se modifique cualquier resultado si le es injustamente adverso [9].

El objetivo de la transparencia es dar la información adecuada a cada destinatario para que pueda comprender y desarrollar confianza en el sistema y, en última instancia, en su propio gobierno o empresa que le da un servicio. La transparencia puede ayudar a las personas a comprender cómo se lleva a cabo cada etapa y revela información si se han establecido las garantías adecuadas, tales como medidas de seguridad o imparcialidad.

Por su parte, la explicabilidad es la capacidad de hacer comprensibles los resultados de los sistemas de IA. Esto también incluye comprender la entrada, la salida, el funcionamiento y la contribución de cada elemento de construcción algorítmica al producto final. Los resultados y los subprocesos también deben ser comprensibles y rastreables [9].

8. Responsabilidad y rendición de cuentas

La UNESCO enfatiza que la responsabilidad debe estar en conformidad con la legislación internacional y nacional de cada país, en particular en materia de derechos humanos. Los actores de la IA deben asumir siempre su deber ético y responsabilidad de las decisiones y acciones que se basen de algún modo en un sistema de IA. Para garantizar la responsabilidad deben diseñarse procesos adecuados de supervisión, de impacto y de evaluación a través de auditorías de cada proceso [1]. Este enfoque multifacético busca reforzar el marco ético en materia de rendición de cuentas y promueve una cultura de transparencia, confianza e implantación responsable de la IA.

9. Sensibilización y educación

La UNESCO aconseja que, para garantizar la eficacia de las políticas públicas, es imperativo que los gobiernos, las organizaciones no gubernamentales, el mundo académico, los medios de comunicación, los líderes comunitarios, las asociaciones civiles y el sector privado colaboren en la promoción de la alfabetización mediática, el compromiso cívico y la formación en competencias digitales. Estos esfuerzos deben tener en cuenta la diversidad cultural y lingüística existente. Se calcula que en la actualidad hay unas doscientas lenguas oficiales reconocidas en el mundo [1].

Además, debe hacerse un esfuerzo concertado para cultivar una cultura de IA ética en todos los sectores, incluidos el académico, la investigación, el gobierno y la

industria. Para ello, los sistemas de IA deben diseñarse teniendo en cuenta consideraciones prácticas responsables, como la transparencia, la rendición de cuentas y el respeto de los derechos humanos. Establecer mecanismos de supervisión y evaluación continuas de los sistemas de IA para garantizar su conformidad con las normas y estándares éticos [10].

10. Gobernanza y colaboración adaptativas

La UNESCO recomienda que las grandes empresas multinacionales de uso masivo de información deben respetar tanto la soberanía nacional como el derecho internacional. Los países son libres de regular los datos creados en su territorio o que pasan por él, y de adoptar medidas para una vigilancia eficaz del uso de la información de sus ciudadanos, incluida la protección y el respeto del derecho a la intimidad, y otras normas de derecho fundamental [1].

Para garantizar la distribución equitativa de los beneficios y la promoción del desarrollo sostenible en IA, es crucial contar con la participación de todas las entidades pertinentes, incluidos los gobiernos, las organizaciones no gubernamentales, la sociedad civil, el mundo académico, los medios de comunicación, los educadores y los responsables de la toma de decisiones de los sectores público y privado. La colaboración entre las partes involucradas es necesaria para facilitar la adopción de normas abiertas e interoperables [11].

El organismo multilateral también recomienda tomar medidas para tener en cuenta los cambios tecnológicos, la creación de nuevos grupos y la participación significativa de comunidades e individuos marginados, así como, el respeto a la autogestión de sus datos. La alfabetización digital puede ser posible en todos los ámbitos, sin traicionar costumbres o cultura [12].

En diciembre de 2022 se realizó el “Primer foro global sobre la ética de la inteligencia artificial”, con el tema “Asegurando la inclusión en el mundo de la IA” y fue patrocinado por UNESCO y organizado por la República Checa en el marco de la presidencia del Consejo de la Unión Europea. Este foro marcó nuevas pautas en la construcción de una coalición internacional para garantizar el desarrollo y el uso éticos de la inteligencia artificial en todo el mundo.

Conclusiones parciales

Hasta aquí los diez principios y valores que la UNESCO recomienda para la creación, desarrollo, uso y aplicación de los sistemas de IA firmados por 193 países. Debe considerarse su carácter propositivo y que, desafortunadamente, algunos países no son signatarios del acuerdo internacional y, por tanto, no se sienten obligados a acatar estas recomendaciones.

Sin embargo, podría decirse, que es lo más cercano a un Acuerdo Internacional avalado por la Organización de las Naciones Unidas en materia de uso responsable de sistemas de IA. La Recomendación es exhaustiva y abarcan una amplia gama de temas. Las recomendaciones específicas piden que el desarrollo y el uso de las tecnologías de IA respeten la dignidad, los derechos y el bienestar humanos, así como la protección del medio ambiente y la justicia social. Esto puede servir de guía a todos los países y empresas del sector.

Pide que la IA aumente y potencie las capacidades humanas, no que las disminuya. Una IA que sirva al ser humano y que mejore y amplifique sus capacidades, en lugar de sustituirlas.

La recomendación trata de reconocer los beneficios que la IA ofrece a la sociedad y reducir los peligros que ésta puede significar. Al abordar cuestiones de transparencia, rendición de cuentas, privacidad y ofrecer aspectos políticos orientados a la acción sobre gobernanza de datos, educación, cultura, trabajo, salud y economía, garantiza que las transformaciones digitales promuevan los derechos humanos y ayuden a alcanzar los Objetivos de Desarrollo Sostenible de la ONU.

Finalmente, el punto central es que sienta las bases de los instrumentos normativos futuros que podrían coadyuvar en su ejecución y que las medidas adecuadas para garantizar que la ética se aplique en los hechos. Se espera que la Recomendación ayude a los países y empresas a mejorar sus marcos ético-normativos desde una visión cosmopolita.

Referencias

[1] UNESCO, “Recommendation on the ethics of artificial intelligence,” 2021, acceso feb. 2023. [En línea]. Disponible: https://bsu.buap.mx/b2m

[2] B. Shneiderman, “Human-centered artificial intelligence: Reliable, safe & trustworthy,” Int. J. Hum. Comput. Interact., vol. 36, no. 6, pp. 495-504, 2020. arXiv:2002.04087v2.

[3] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo y L. Floridi, "The ethics of algorithms: key problems and solutions," *AI & Soc.*, vol. 37, no. 1, pp. 215-230, Feb. 2022. doi: 10.1007/s00146-021-01154-8.

[4] P. Boddington, "Normative Modes: Codes and Standards," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford: Oxford University Press, 2020, pp. 124-140.

[5] J. Cowls, A. Tsamados, M. Taddeo y L. Floridi, "The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations," *AI & Soc.*, vol. 38, pp. 283-307, 2023. doi:10.1007/s00146-021-01228-6.

[6] C. Véliz, "Privacy Is Power: Why and How You Should Take Back Control of Your." NY: Penguin Random House, 2022.

[7] AlgorithmWatch, "AlgorithmWatch is a non-profit research and advocacy organization that is committed to watch, unpack and analyze automated decision-making (ADM) systems and their impact on society," Algorithmwatch.org. Acceso feb. 2023. [En línea]. Disponible: https://algorithmwatch.org/

[8] A. Smith, "Using Artificial Intelligence and Algorithms," FTC.gov. Acceso feb. 2023. [En línea]. Disponible: https://www.ftc.gov/business-guidance/blog/2020/04/using-artificial-intelligence-and-algorithms.

[9] N. Diakopoulos, "Transparency. Accountability, Transparency, and Algorithms," in *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford, UK: Oxford University Press, 2020, pp. 197-213.

[10] D. Peters, P. H. Kuss, T. Calders y F. Schram, "Responsible AI—two frameworks for ethical design practice," *IEEE Trans. on Tech. and Soc.*, vol. 1, no. 1, pp. 34-47, 2020. Disponible: https://bsu.buap.mx/ciQ

[11] M. Taddeo y L. Floridi, "How AI Can Be a Force for Good – An Ethical Framework to Harness the Potential of AI While Keeping Humans in Control," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed. Cham: Springer International Publishing, 2021, pp. 91-96.

[12] T. Powers y J. Ganascia, "The Ethics of the Ethics of AI," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Editors, Oxford University Press, 2020, pp. 26-51.

RECOMENDACIÓN DEL CONSEJO SOBRE INTELIGENCIA ARTIFICIAL DE LA OCDE: DESIGUALDAD E INCLUSIÓN

Introducción

La “Recomendación del Consejo de la OCDE sobre inteligencia artificial” es un conjunto de directrices y principios para el desarrollo y despliegue de sistemas de IA alineados con los valores y derechos humanos, la transparencia y la rendición de cuentas. Fue adoptada por el Consejo de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) en mayo de 2019 e incluye cinco principios clave para la gobernanza de la IA. Ésta debe beneficiar a las personas y al planeta impulsando el crecimiento inclusivo, el desarrollo sostenible y el bienestar; los sistemas de IA deben respetar el estado de derecho, los derechos humanos, los valores democráticos y la diversidad; también deben ser transparentes para garantizar que las personas comprendan los resultados y puedan cuestionarlos; los sistemas deben ser sólidos, seguros y protegidos durante todo su ciclo de vida; y las partes interesadas deben colaborar para maximizar los beneficios de la IA y minimizar sus riesgos. Se concluye aquí que los Principios de IA de la OCDE han sido criticados por no ser vinculantes y carecer de un mecanismo formal de aplicación. Tampoco abordan preocupaciones sociales amplias, como el desplazamiento de puestos de trabajo y la concentración de poder en unas pocas empresas. Además, es posible que los principios no tengan en cuenta los retos a los que se enfrentan los países en desarrollo y estén demasiado centrados en los intereses de las economías avanzadas. Sin embargo, la Recomendación de la OCDE establece una serie de principios ampliamente aceptados y respaldados tanto por los miembros, como por los que no lo son. Entre los valores figuran la transparencia, la solidez, la seguridad, la responsabilidad, la justicia, la equidad y la determinación humana. Este conjunto compartido de principios proporciona una base sólida para el desarrollo responsable de la IA, más allá de las fronteras de los “países ricos”.

Recomendación del consejo de la OCDE sobre inteligencia artificial

El Consejo de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), a nivel ministerial, aceptó la Directriz sobre inteligencia artificial en mayo de 2019 a propuesta del Comité de Política de Economía Digital (CDEP por sus siglas en inglés) [1]. La "Recomendación del consejo de la OCDE sobre inteligencia artificial" (Recommendation of the Council on OECD Legal Instruments Artificial Intelligence - OECD-LEGAL-0449) [2] es un documento que promueve la administración responsable de la IA, al tiempo que trata de garantizar el respeto de los derechos humanos y los ideales democráticos, con el fin de promover la innovación y confianza en la IA.

La Recomendación de OCDE, se centran en cuestiones específicas de la IA y establece una norma aplicable y flexible, para que no quede pronto obsoleta, complementando las normas existentes de la organización en áreas como la privacidad, la gestión de riesgos de seguridad digital y la conducta empresarial responsable.

La OCDE, que tiene su sede en París y está formada por 38 Estados con el objetivo de coordinar sus políticas económicas y sociales, también se conoce como el "club de los países ricos", debido a que entre sus agremiados se concentra más del 60% del PIB nominal mundial [3]. Sin embargo, hay que señalar que países miembros como Turquía, Colombia, Chile y México aún enfrentan retos económicos importantes como para ser considerados "países ricos" [4].

La "Recomendación del consejo sobre inteligencia artificial" es un documento no vinculante que fue aprobado y firmado el 21 de mayo de 2019 por los 38 países miembros, a los que se adhirieron ese mismo día, Argentina, Brasil, Perú y Rumania (no miembros); posteriormente, lo hicieron otros no miembros, como Ucrania (29 de octubre de 2019), Malta (19 de diciembre de 2019), Egipto y Singapur (30 de marzo de 2021). Estas directrices están en concordancia con otras de la propia organización en ámbitos como la protección de datos, la gestión de riesgos para la seguridad digital y las prácticas empresariales éticas [5].

Así, el Comité de expertos en inteligencia artificial de la OCDE publicó en mayo de 2019 este conjunto de principios éticos que fue adoptado por los países miembros y adherentes.

1. Crecimiento inclusivo, desarrollo sostenible y bienestar

El principio destaca la capacidad de la IA fiable para promover los objetivos de desarrollo mundial, el progreso y la prosperidad generales para todos los involucrados alrededor del orbe.

La OCDE ha hecho un llamamiento a las naciones más avanzadas tecnológicamente, instándolas a asumir un deber internacional de solidaridad con las naciones menos avanzadas tecnológicamente. Esto sirve para garantizar que los beneficios de la IA se distribuyan de manera que se fomente el acceso y la participación, contribuyendo a un orden mundial más equitativo con respecto a la información, la educación, la cultura, la investigación y la estabilidad de las naciones [2].

Para lograr este objetivo, es esencial establecer un marco de cooperación y colaboración internacional. Dicho marco implicaría el intercambio de conocimientos y experiencia, la provisión de recursos financieros y tecnológicos y el compromiso activo de todas las partes interesadas. Si se construye una comunidad mundial integradora y cooperativa, se puede crear un entorno en el que la mayoría disfrute de los beneficios de la IA y en el que el progreso tecnológico se aproveche para consolidar el desarrollo social [6].

2. Valores y equidad centrados en el ser humano

Para garantizar una sociedad justa y equitativa, la OCDE considera que los sistemas de IA deben crearse de manera que respeten el estado de derecho, los principios democráticos y la diversidad.

A lo largo de la vida útil de los sistemas de inteligencia artificial, los agentes deben respetar la ley, los derechos humanos y los principios democráticos. Por ejemplo, la diversidad, la equidad, la justicia social, la no discriminación, la autonomía, privacidad, protección de datos y los derechos laborales reconocidos. Para lo anterior, los actores de IA deben implementar garantías y métodos que sean apropiados y en línea con las mejores prácticas, como la prioridad del juicio humano y la autodeterminación [2].

3. Transparencia y explicabilidad

Este principio sirve para que los usuarios sepan cuándo están interactuando con sistemas de IA y puedan cuestionar los resultados; este principio enfatiza la divulgación responsable para que cualquiera entienda los mecanismos de decisión automática [2].

La OCDE destaca la importancia de que los agentes de la IA se comprometan a proporcionar datos pertinentes que se ajusten al estado actual de la tecnología y sean adecuados para la situación específica. Por ejemplo, esto incluye aumentar la responsabilidad entre las partes interesadas sobre sus interacciones con los sistemas de IA, tanto en el lugar de trabajo como fuera de él. También implica fomentar una comprensión general de los sistemas de IA para que los afectados puedan comprender los resultados, sobre todo si son desfavorables. Esta comprensión permite a las personas cuestionar las decisiones basándose en información transparente y comprensible sobre las variables implicadas y el razonamiento que subyace a las predicciones, recomendaciones y decisiones [7].

4. Robustez, seguridad y protección

Este principio se refiere a que, a lo largo de su vida útil, los sistemas de IA deben funcionar de manera fiable, segura y protegida de los posibles riesgos, que siempre deben ser evaluados y abordados oportunamente [2].

La OCDE considera que los sistemas de IA deben ser fiables, seguros y protegidos durante todo su ciclo de vida para que puedan funcionar correctamente y no supongan un peligro excesivo para la seguridad en situaciones de uso normal y previsible [2].

Para permitir el análisis de los resultados de los sistemas de IA y las respuestas a las consultas es crucial que las partes interesadas garanticen la trazabilidad a lo largo del ciclo de vida del sistema. Esto implica hacer un seguimiento de los conjuntos de datos, los procesos y las decisiones, y garantizar que las respuestas sean adecuadas al contexto específico y coherentes con la tecnología y las prácticas más recientes.

La trazabilidad es un concepto clave en este contexto, ya que permite el seguimiento de todos los procesos desde los datos de entrada iniciales hasta los resultados de salida. Al garantizar el seguimiento y registro de toda la información

pertinente, las partes interesadas pueden identificar y abordar cualquier problema o error que pueda surgir [8].

Además, las partes deben tener en cuenta la importancia de cada contexto, las circunstancias particulares en las que opera el sistema de IA a la hora de determinar la respuesta adecuada, e incluir factores como la diversidad cultural y lingüística, así como los marcos jurídicos y normativos.

Garantizar la trazabilidad es esencial para promover la responsabilidad y la transparencia en los sistemas de IA. Mediante el seguimiento y el análisis de todos los conjuntos de datos, procesos y decisiones pertinentes, los involucrados pueden contribuir a garantizar que los sistemas de IA funcionen de forma ética y responsable, en consonancia con las mejores prácticas y tecnologías [9]. Para gestionar los riesgos vinculados a los sistemas de IA, como la privacidad, la seguridad digital, y la parcialidad, la OCDE afirma que se debe aplicar continuamente una estrategia sistemática de gestión de riesgos en cada fase del ciclo de vida del sistema, en función del contexto y la capacidad de actuación [2].

5. Responsabilidad

El correcto funcionamiento de la IA, de acuerdo con las directrices de la OCDE, debe ser responsabilidad de las organizaciones y personas que los gestionan, desarrollan y despliegan, en función de sus compromisos, del contexto, y de acuerdo con la situación imperante. Los actores de la IA deben ser responsables del correcto funcionamiento de los sistemas y no soslayar el hecho de que, detrás de todo sistema, finalmente son personas quienes los administran [2].

En octubre de 2021, una nueva Convención denominada “Poniendo en práctica los principios de IA de la OCDE: avances y perspectivas futuras” [10] revisó los principios formulados dos años antes para validar su actualidad, y consideraron cinco recomendaciones más para los gobiernos en materia de IA:

- Invertir en investigación y desarrollo.
- Fomentar un ecosistema digital propicio.
- Dar forma a un entorno político adecuado sobre IA.
- Desarrollo de capacidades para un nuevo mercado laboral y,
- Cooperación internacional para una IA confiable.

Las prioridades, por tanto, además de dar financiamiento a las instituciones y proyectos de investigación y desarrollo sobre IA, abordan los problemas sociales y fomentan la adopción de la IA por parte de las empresas [11].

Asimismo, buscan impulsar el diálogo social inclusivo, así como dotar a la población de las competencias necesarias para utilizar la IA y promover una transición justa de los trabajadores al nuevo mercado laboral, con una alta tendencia a la automatización de los procesos [12].

Estas fueron algunas de las prioridades estratégicas y políticas establecidas por la OCDE en 2021. La Recomendación también impulsó a la OCDE a crear un "Observatorio de políticas de IA" [16] para ayudar a poner en práctica las directrices y vigilarlas.

Conclusiones parciales

Para la OCDE la IA debe promover el crecimiento equitativo, el desarrollo sostenible y el bienestar general en beneficio de las personas y el medio ambiente; para garantizar una sociedad justa y equitativa, los sistemas de IA deben construirse de forma que respeten el estado de derecho, los derechos humanos, los principios democráticos, la diversidad e incorporen las protecciones adecuadas para permitir la intervención humana cuando sea necesario [13].

También garantizar que las personas sean conscientes de cuándo están interactuando con los sistemas de IA y tengan la capacidad de impugnar los resultados de esas interacciones. Así, los sistemas de IA deben guiarse por la transparencia y la ejecución responsable.

El principal objetivo de la "Recomendación del consejo de la OCDE" es fomentar la innovación y la confianza en la IA, al tiempo que se defienden los derechos humanos y los principios democráticos. La Recomendación trata de alcanzar un consenso sobre la importancia y los conceptos rectores de una tecnología fiable y apoya una IA propositiva pero segura [14].

Al igual que otras propuestas, las críticas de este instrumento han sido que sus principios no son vinculantes, lo que significa que no existe un mecanismo formal para hacerlos cumplir. Los investigadores sostienen que esta falta de aplicación

hace que los principios no tengan fuerza legal y quede a consideración de los desarrolladores en las empresas y los gobiernos acatarlas, aun siendo signatarios de la OCDE [7].

Las críticas se han centrado en que, dado que los principios son amplios, están abiertos a la interpretación, lo que puede dificultar su aplicación en la práctica. Los principios se concentran en la gobernanza de la IA, pero no abordan preocupaciones más amplias relacionadas con el impacto que la inteligencia artificial pueda tener en la sociedad; por ejemplo, se argumenta que no abordan específicamente cuestiones como el desplazamiento de puestos de trabajo, el sesgo algorítmico o la concentración de poder en manos de unas pocas empresas y la falta de inclusividad [15]. Esto es que no tienen en cuenta los retos y contextos específicos de los países en desarrollo y pueden estar demasiado cerca de los intereses de las economías avanzadas [15].

En conclusión, la "Recomendación del consejo de la OCDE sobre inteligencia artificial" [2] subraya la urgente necesidad de abordar los dilemas del poder y la inclusión para garantizar que los beneficios de la IA sean compartidos por todos, al tiempo que se minimizan los riesgos potenciales y los impactos negativos. Corresponde a los responsables políticos, a las partes interesadas y a la sociedad civil trabajar juntos hacia un enfoque de la IA centrado en el ser humano que promueva un desarrollo y un despliegue responsables, transparentes e inclusivos de esta poderosa tecnología.

Si bien la OCDE puede representar los intereses de sus miembros, lo cierto es que este breve documento, sin ser legalmente normativo, también es una propuesta seria, directa y comprometida por sentar las bases para una ley de aspiraciones universalistas para el bien común.

Referencias

[1] OCDE, "Cuarenta y dos países adoptan los Principios de la OCDE sobre Inteligencia Artificial," OECD.org, 2019. Acceso mar. 2023. [En línea]. Disponible: https://bsu.buap.mx/b2r

[2] OCDE, "Recommendation of the Council on Artificial Intelligence," OECD.org 2019. OECD.org, 2019. Acceso mar. 2023. [En línea]. Disponible https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

[3] Bank, W., "GDP growth OECD members," 2021. Acceso mar. 2023. [En línea]. Disponible: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=OE

[4] J. Clifton y D. Díaz‐Fuentes, “From ‘Club of the Rich’ to ‘Globalisation à la carte’? Evaluating Reform at the OECD,” *Global Policy*, vol. 2, no. 3, pp. 300-311, 2011, doi: 10.1111/j.1758-5899.2011.00103.x.

[5] OCDE, “Papers,” OECD.org. Acceso feb. 2023. [En línea]. Disponible: https://www.oecd-ilibrary.org/papers

[6] A. Korinek, “Integrating Ethical Values and Economic Value to Steer Progress in Artificial Intelligence,” in *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford, UK: Oxford University Press, 2020, pp. 475–491.

[7] AlgorithmWatch, “AlgorithmWatch is a non-profit research and advocacy organization that is committed to watch, unpack and analyze automated decision-making (ADM) systems and their impact on society,” 2022. Acceso feb. 2023. [En línea]. Disponible: https://algorithmwatch.org/

[8] S. Russell y P. Norvig, “Philosophy, ethics, and safety of AI,” en *Artificial Intelligence: A Modern Approach*, Londres: Pearson, 2022, pp. 1032-1062.

[9] R. Blackman, “Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI.” Harvard, MA, USA: Harvard Business Review Press, 2022.

[10] OCDE, “Putting the OECD AI Principles into practice: progress and future perspectives,” OECD.ai, 2021. Acceso feb. 2023. [En línea]. Disponible: https://oecd.ai/en/mcm

[11] K. Yeung, “Recommendation of the Council on Artificial Intelligence (OECD),” *Intern. Legal Materials*, vol. 59, no. 1, pp. 27-34, 2020.

[12] OCDE, “Employment database - Unemployment indicators,” OECD.org. Acceso feb. 2023. [En línea]. Disponible: https://t.ly/CEO76

[13] C. Burr, M. Taddeo, y L. Floridi, “The Ethics of Digital Well-Being: A Thematic Review,” *Sci. Eng. Ethics*, vol. 26, no. 4, pp. 2313-2343, 2020, doi: 10.1007/s11948-020-00175-8

[14] S. Milano, M. Taddeo, y L. Floridi, “Recommender systems and their ethical challenges,” *AI & Soc.*, vol. 35, no. 4, pp. 957-967, 2020, doi: 10.1007/s00146-020-00950-y

[15] Nature, “International AI ethics panel must be independent,” Nature.com. ago. 2019. Acceso feb. 2023. [En línea]. Disponible: https://www.nature.com/articles/d41586-019-02491-x

[16] OECD, “OECD.AI Policy Observatory”. Acceso feb. 2023. [En línea]. Disponible: https://oecd.ai/en/

PROPUESTA AXIOLÓGICA DE LA UNIÓN EUROPEA EN INTELIGENCIA ARTIFICIAL: "DIRECTRICES ÉTICAS PARA UNA IA CONFIABLE"

Introducción

Las "Directrices éticas para una IA digna de confianza" de la Unión Europea fueron elaboradas por un grupo independiente de más de cincuenta expertos de alto nivel sobre inteligencia artificial de diciembre de 2018 a abril de 2019, nombrados por la Comisión Europea. Las directrices se basan en los derechos fundamentales consagrados en la "Carta de los derechos fundamentales de la UE", así como en una consulta pública realizada. Las Directrices pretenden garantizar que la IA se desarrolle y utilice, de forma fiable y respetuosa con los derechos inalienables del ser humano. Establecen siete requisitos clave para una IA digna de confianza, entre ellos el respeto a la autonomía de la persona, la prevención de daños, la transparencia, la diversidad y la equidad. Algunos analistas han argumentado que las directrices son demasiado amplias y carecen de especificidad, lo que dificulta su aplicación en la práctica. Al igual que otras directrices para IA, aquí se señala que no son jurídicamente vinculantes y, por tanto, carecen de fuerza para garantizar su cumplimiento por parte de empresas y gobiernos. Sin embargo, las Directrices se consideran un paso importante y firme hacia la creación de un marco para el desarrollo y despliegue responsables de la IA dentro y fuera del continente europeo.

Directrices éticas para una IA digna de confianza de la Unión Europea

En abril de 2019, un grupo independiente de más de cincuenta expertos de alto nivel sobre inteligencia artificial (High-Level Expert Group on AI - AI HLEG), convocados por el Consejo de Ministros de la Unión Europea, publicó el documento "Directrices éticas para una IA digna de confianza" (Ethics Guidelines for Trustworthy AI) [1], que

incluye un conjunto de principios básicos y responsabilidades para guiar el desarrollo y el uso de la IA en la Unión Europea, pero que, por supuesto, aspira a tener una validez universal.

La lista de los expertos convocados en orden alfabético es la siguiente: Aimee Van Wynsberghe, Andrea Renda, Barry O'Sullivan, Catelijne Muller, Cecile Wendling, Cecilia Bonefeld-Dahl, Chiara Giovannini, Christoph Peylo, Cristina San José, Elisabeth Ling, Eric Hilgendorf, Fanny Hidvegi, Francesca Rossi, Françoise Soulié Fogelman, Fredrik Heintz, George Sharkov, Gry Hasselbalch, Ieva Martinkenaite, Iris Plöger, Jaan Tallinn, Jakob Uszkoreit, Jean-Francois Gagné, Joanna Goodey, Karen Yeung, Klaus Höckner, Leo Kärkkäinen, Loubna Bouarfa, Luciano Floridi, Mária Bieliková, Mari-Noëlle Jégo-Laveissière, Mark Coeckelbergh, Markus Noga, Nicolas Petit, Pekka Ala-Pietilä, Philipp Slusallek, Pierre Lucas Orgalim, Raja Chatila, Robert Kroplewski, Sabine Theresia Köszegi, Sami Haddadin, Saskia Steinacker, Stefano Quintarelli, Stéphan Brunessaux, Thiébaut Weber, Thierry Tingaud, Thomas Metzinger, Urs Bergmann, Ursula Pachl, Virginia Dignum, Wilhelm Bauer y Yann Bonnet [2].

En el comunicado final, el grupo de expertos sostuvo que la estrategia del documento tiene como objetivo aprovechar el potencial de la IA para mejorar la vida de las personas y la economía europea, mientras se asegure de que la tecnología se desarrolle y se use de manera responsable y ética. La estrategia incluye una serie de medidas para fomentar la investigación y el desarrollo en el campo de la IA, establecer normas éticas y garantizar la protección de los derechos humanos y la privacidad.

El primer capítulo del texto establece las bases de una IA fiable introduciendo su enfoque fundamental basado en los derechos. Esboza los principios éticos que deben respetarse para garantizar una IA confiable y ética, proporcionando una descripción de cada principio. Los cuatro principios que se proponen son los siguientes:

1. El principio del respeto de la autonomía humana

La Unión Europea (UE) ha identificado la necesidad de defender los derechos fundamentales como la libertad y la autonomía individuales durante las interacciones con los sistemas de IA. Es crucial, dicen, que los individuos puedan mantener el control sobre sí mismos y participar en procesos democráticos sin ningún tipo de coacción, manipulación o dirección injusta por parte de los sistemas de inteligencia artificial [1].

Es importante prestar mucha atención a las circunstancias que tienen un impacto significativo en las personas más susceptibles de sufrir daños, como los niños, las personas con discapacidad y otras que se han enfrentado a desventajas o corren riesgo de exclusión. Además, se debe ser consciente de las situaciones en las que existe una distribución desigual del poder o de la información, como las que pueden darse entre gobiernos y ciudadanos, empresarios y trabajadores o entre empresas y consumidores [3] [4].

2. El principio de prevención del daño

Los sistemas de IA no deben dañar ni afectar negativamente a los seres humanos de ninguna manera. Esto incluye proteger la dignidad humana y el bienestar físico y mental, por lo que es importante garantizar que los sistemas y sus entornos operativos sean seguros, técnicamente sólidos y éticos [1]. Del mismo modo es importante asegurarse de que sean técnicamente fuertes y no susceptibles al uso malicioso. Las personas que son vulnerables debido a los desequilibrios de poder o la asimetría de la información deben recibir más atención en el desarrollo, el despliegue y el uso de los sistemas de IA [5] [6]. Además, se debe prestar especial atención a los escenarios en los que los sistemas de IA pueden causar o empeorar las consecuencias negativas. Todos estos elementos pueden ser aglutinados en la idea general del principio de prevención del daño.

3. El principio de equidad

La implantación, ejecución y utilización equitativa de los sistemas de IA son imperativas. A pesar de las diversas definiciones de equidad, se afirma que comprende dos aspectos esenciales: procedimental y sustantivo [1].

La dimensión sustantiva de la equidad implica una dedicación a lograr la paridad en la asignación justa de ventajas y desventajas, así como a salvaguardar a las personas y los grupos de los prejuicios, la discriminación y el etiquetado injusto [7]. Si se da prioridad a evitar los prejuicios, los sistemas de IA pueden contribuir al avance de la sociedad. Además, también debería fomentarse la igualdad de oportunidades en la educación, el acceso a los bienes, los servicios y la tecnología [5].

4. El principio de explicabilidad

Para establecer y mantener la confianza de los usuarios en los sistemas de IA, es crucial garantizar la transparencia de los procesos, la comunicación abierta de las capacidades y objetivos de los sistemas de IA para poder explicar las decisiones a

los afectados directa e indirectamente. Si no se facilita esta información, la decisión no podrá ser impugnada [1].

En algunos casos, puede resultar difícil explicar el razonamiento que subyace al resultado de un sistema de IA debido a su uso de algoritmos de caja negra [8], como ya se ha dicho, se denomina así a los algoritmos que se conocen sus datos de entrada, los de salida, pero no cómo llegó a sus resultados. Por lo tanto, pueden ser necesarios métodos alternativos de explicabilidad, como la trazabilidad, la auditabilidad y la comunicación transparente de las capacidades del sistema [9]. Sin embargo, estas medidas no deben comprometer la seguridad general [10]. El alcance de la explicación necesaria depende en gran medida del contexto y de la gravedad de las consecuencias si los resultados son incorrectos o inexactos [11].

El segundo capítulo de las Directrices toma los principios éticos introducidos en el primer capítulo y los convierte en siete requisitos esenciales que los sistemas de IA deben satisfacer a lo largo de toda su vida útil para lograr que sea digna de confianza. El capítulo sugiere enfoques técnicos y no técnicos que pueden emplearse para cumplir estos requisitos. Los requisitos para garantizar confianza en los sistemas de IA, según el grupo de expertos de la UE son:

1. Agencia humana y supervisión

Los sistemas de IA deben respetar la autonomía humana y apoyar la toma de decisiones, promover los derechos fundamentales minimizando los impactos negativos. Los usuarios deben tener los conocimientos y las herramientas necesarias para interactuar con los sistemas de IA y tomar decisiones debidamente informadas, por lo que los sistemas de IA no deben manipular ni condicionar injustamente el comportamiento humano, por ejemplo, en su decisión de voto en un ambiente democrático.

La supervisión es crucial para evitar efectos adversos en la autonomía individual, esta puede lograrse mediante mecanismos de gobernanza como la participación humana, el control y el mando. Las autoridades deben poder supervisar los sistemas de IA, y pueden ser necesarios mecanismos de control en función del alcance y los riesgos del sistema [1] [12].

2. Solidez técnica y seguridad

Los sistemas de IA también deben incorporar resistencia a los ataques y medidas de seguridad, tener planes alternativos para posibles fallos y garantizar la

fortaleza, precisión, fiabilidad y reproducibilidad generales [1]. La solidez técnica es un elemento crítico para establecer una IA digna de confianza y está estrechamente relacionada con el principio de prevención de daños. Para garantizar una mejor comprensión, los sistemas de IA deben diseñarse con un enfoque proactivo de la gestión de riesgos. Deben comportarse sistemáticamente según lo previsto, minimizando la aparición de daños no deseados e imprevistos y evitando al mismo tiempo cualquier forma de daño inaceptable [13]. Esto también debe tener en cuenta los cambios en el entorno operativo del sistema y la presencia de otros agentes, incluidos los agentes malintencionados humanos y artificiales.

3. Gestión de la privacidad y de los datos

Respeto a la privacidad, el mantenimiento de la calidad e integridad de los datos y la garantía de un acceso adecuado a los mismos [1]. El principio de prevención del daño está intrínsecamente relacionado con el derecho básico a la intimidad, que se ve afectado significativamente por los sistemas de IA abusivos [14]. Una administración y gestión adecuadas de los datos es esencial para salvaguardar la privacidad, y esto abarca la calidad de los datos, la integridad, la pertinencia para el entorno operativo del sistema de IA, los protocolos de acceso y la capacidad de procesar información garantizando, al mismo tiempo, la protección de los datos personales [15]. Es peligroso que los datos de una persona sean publicados sin su consentimiento porque existen en la red muchas fuerzas dedicadas al robo de identidad, el fraude y la delincuencia, tanto en la web profunda (ver glosario) como en la web visible [16].

4. Transparencia

La trazabilidad, la explicabilidad y la comunicación eficaz son también elementos esenciales de una IA confiable [1]. La trazabilidad en los sistemas de IA se refiere a documentar los conjuntos de datos, procesos y algoritmos utilizados en la toma de decisiones. El principio de explicabilidad está estrechamente ligado a hacer transparentes los aspectos relevantes como los datos utilizados, y los modelos algorítmicos asociados [17]. En cuanto a la comunicación, los sistemas de IA deben poder distinguirse de los humanos, y los usuarios tienen derecho a saber cuándo están en contacto con una IA y tener la opción de elegir si quieren interactuar con un humano. Los usuarios finales deben ser informados de las capacidades y limitaciones del sistema de IA. La información facilitada debe adaptarse al caso de uso específico e incluir detalles sobre la precisión y alcances del sistema.

5. Diversidad, no discriminación y equidad

Necesidad de evitar sesgos injustos, accesibilidad, diseño universal y participación de las partes interesadas, conforman este principio [1]. Para establecer una IA digna de confianza, es crucial dar prioridad a la inclusión y la diversidad en cada fase del desarrollo de un sistema de IA. Esto significa tener en cuenta e implicar activamente a todas las partes interesadas que se vean afectadas por el sistema, por ejemplo, quienes tengan alguna discapacidad, garantizar el acceso mediante procesos de diseño y tratar a todos con equidad. También se considera que, en un sistema de automatización del trabajo, los empleados deben estar involucrados e informados del desarrollo de los nuevos mecanismos [17].

6. Bienestar social y ambiental

Este precepto tiene tres dimensiones: sustentabilidad, impacto social y democracia.

Promover la sostenibilidad y el respeto al medio ambiente, tener en cuenta el impacto social de los sistemas de IA y fomentar los valores de la sociedad y la democracia [1]. Para garantizar los principios de equidad y prevención de daños, es esencial reconocer a la sociedad, a otros seres sensibles y al medio ambiente como partes involucradas a lo largo de todo el ciclo de vida del sistema de IA. Esto incluye fomentar la sostenibilidad y la responsabilidad ecológica en los sistemas, promover la investigación de soluciones que aborden retos globales como los Objetivos de Desarrollo Sostenible, y utilizar los sistemas de IA para la mejora de todos los seres humanos, incluidas las generaciones futuras [18].

La exposición a los sistemas de IA en diferentes ámbitos de la vida, como la educación, el trabajo, las interacciones sociales y las actividades de ocio, tiene el potencial de influir en la forma en que los individuos perciben las acciones sociales y afectar a sus relaciones. Aunque los sistemas de IA pueden emplearse para mejorar las habilidades interpersonales, también pueden contribuir a su deterioro, lo que puede repercutir en el bienestar físico y mental de las personas [1].

A la hora de evaluar el impacto de la inteligencia artificial, es crucial evaluar sus efectos sobre los individuos, las instituciones, la democracia y la sociedad en general. Deben realizarse estudios detallados para comprender las repercusiones de los sistemas de IA, en particular en relación con el proceso democrático, que abarca la toma de decisiones políticas y los contextos electorales [1].

7. Rendición de cuentas

Este precepto tiene cuatro dimensiones en el documento: auditabilidad, minimización de efectos negativos, equilibrios y compensaciones.

La auditabilidad se refiere a la capacidad que se tiene de evaluar los algoritmos, los datos de entrada, los procesos y los resultados de su diseño. Segundo, minimizar y notificar las repercusiones negativas es tener en cuenta las ventajas y desventajas, así como ofrecer mecanismos de reparación en los usos de sistemas de IA [1]. Tercero, buscar que exista equilibrio entre intereses y valores subyacentes al sistema de IA. Cuarto, la compensación exige el establecimiento de mecanismos que garanticen que, en caso de efectos adversos, los afectados sean retribuidos [19].

Derivado de un segundo documento denominado: "Recomendaciones de política e inversión para una inteligencia artificial confiable" [Policy and investment recommendations for trustworthy Artificial Intelligence] [20] el AI HLEG lanzó en 2019 las siguientes cuatro consideraciones:

1. Inversiones en investigación e innovación

La Unión Europea planea destinar fondos significativos a la investigación en inteligencia artificial y a la creación de infraestructuras para apoyar el desarrollo de la tecnología [20]. Estas inversiones contribuyen al avance de las tecnologías de IA, haciéndolas más precisas, eficientes y fiables. Esto permite a los sistemas de IA realizar sus tareas con mayor eficacia y tomar mejores decisiones, lo que en última instancia redundará en mejores resultados para las personas, las organizaciones y la sociedad en su conjunto. La Unión Europea ha asignado 7 600 millones de euros al "Programa Europa Digital", el cual tiene como objetivo cubrir la brecha entre la investigación y la implementación de tecnologías digitales, incluyendo capacidades fundamentales de IA, espacios de datos y bibliotecas de algoritmos de IA [31].

2. Desarrollo de capacidades

La UE pretende fomentar el desarrollo de capacidades en el campo de la IA para asegurarse de que la región tenga una fuerza laboral capacitada y experta en la tecnología [20]. Para ello, la UE se ha comprometido a invertir en capacitación destinada a desarrollar las habilidades y conocimientos necesarios para trabajar con IA [22]. Esto incluye iniciativas para promover la educación, el desarrollo de

habilidades especializadas en IA, como el análisis de datos, el aprendizaje automático y el procesamiento del lenguaje natural [23].

3. Normas éticas claras

La estrategia incluye la creación de normas éticas claras para guiar el desarrollo y la utilización de la IA, a fin de garantizar que se respeten los derechos humanos y la privacidad [20]. Una de las principales ventajas de ello es que se contribuye a garantizar que los sistemas de IA se desarrollen de manera benéfica y en el marco del bien común. Por ejemplo, si un sistema de IA está diseñado para recoger y procesar datos personales, debe hacerse de forma transparente y con el consentimiento de las personas cuyos datos se capturan [24].

Otra ventaja es que, cuando las personas saben que los sistemas de IA se desarrollan y utilizan de forma ética y respetuosa con sus derechos, es más probable que confíen en ellos [25]. También es importante señalar que las normas éticas no deben ser estáticas. A medida que las tecnologías de la IA evolucionan, las normas deben revisarse y actualizarse periódicamente para garantizar su pertinencia y eficacia. Esto requerirá un compromiso continuo con las partes interesadas, incluidos los responsables técnicos, los líderes de la industria, los intelectuales y las organizaciones de la sociedad civil.

4. Cooperación internacional

La UE busca colaborar con otros países y regiones para establecer estándares globales en el campo de la IA y trabajar juntos para solucionar los desafíos más importantes que plantea la tecnología [20]. La inteligencia artificial es una cuestión global que afecta a países y regiones de todo el mundo. Por ello, la cooperación y colaboración internacionales para establecer normas mundiales de desarrollo y uso de la IA se convierte en una prioridad. La Unión Europea (UE) reconoce la importancia de la cooperación en el ámbito de la IA y trata de extender sus alcances a otros países y regiones para hacer frente a los retos que plantea esta tecnología [26].

Uno de los principales beneficios de la cooperación internacional es la capacidad de establecer normas compartidas para el desarrollo y uso de la IA, lo que, finalmente, es la aspiración de casi todos los interesados. Esto es importante porque las tecnologías de IA se están desarrollando y utilizando a nivel global, las compañías operan en todas partes, y disponer de normas comunes puede ayudar a garantizar que estas tecnologías se desarrollen y utilicen de forma ética y respetuosa con

los derechos de las personas sin importar su lugar de residencia. Además, normas transnacionales pueden ayudar a evitar que gobiernos y poderes económicos empresariales se extralimiten negativamente en otros países o regiones menos favorecidas [27].

Otro beneficio de la cooperación internacional en IA es la posibilidad de compartir conocimientos y experiencia. La IA es un campo que evoluciona vertiginosamente, especialmente a partir del año 2023, y los países y regiones pueden beneficiarse de compartir sus conocimientos y experiencias entre sí. Esto puede ayudar a acelerar el desarrollo y garantizar que algunas regiones no se queden atrás en este campo, especialmente América Latina, África y algunas regiones alejadas de Asia [26].

Trabajando juntos, los países y empresas pueden desarrollar estrategias para abordar preocupaciones específicas como el impacto de la IA en el empleo y la economía, y garantizar que los beneficios se distribuyan equitativamente entre la sociedad.

Conclusiones parciales

La iniciativa de la UE considera que los cambios inducidos por la IA en las esferas sociales son evidentes, así como en la estabilidad emocional, económica y profesional de todos los involucrados [1]. Por ejemplo, señala que los empleados están preocupados justificadamente por ser reemplazados por robots, los clientes no están seguros de a quién responsabilizar si un sistema basado en IA los afecta, los empresarios tienen problemas para localizar los recursos y el talento que necesitan para su planta productiva, y la competencia internacional en IA es cada vez más feroz debido a inversiones masivas que están realizando países como Estados Unidos y China [28] [29].

De acuerdo con las “Directrices éticas para una IA digna de confianza”, la tecnología debe concentrarse en el ser humano y promover su uso para ayudar a resolver los mayores desafíos del mundo, como son el combatir el cambio climático, anticipar desastres naturales, curar enfermedades, mejorar la seguridad, reducir la delincuencia y reforzar la ciberseguridad.

Algunas críticas al documento de la Unión Europea, al igual que otras normativas, han sido que las directrices son demasiado amplias y carecen de especificidad, lo que dificulta su aplicación en la práctica [21]. En algunos foros [6] se ha expresado la preocupación por la posibilidad de que las directrices frenen la innovación y

obstaculicen el desarrollo de la IA, al imponer requisitos excesivamente restrictivos a los países europeos, mientras otros, como Rusia y China, no están generando esas barreras.

A pesar de estas críticas, las Directrices se consideran un paso importante hacia la creación de un marco para el desarrollo y despliegue responsables de la IA, y han influido en el despliegue de la política y la regulación en la UE, e incluso fuera de ella.

El punto central de la propuesta axiológica de la Unión Europea sobre inteligencia artificial es establecer un enfoque único que beneficie, potencie y proteja el florecimiento humano individual. Las directrices pretenden garantizar que los sistemas de IA sean dignos de confianza y cumplan los requisitos fundamentales aquí expuestos.

Referencias

[1] Unión Europea, "Ethics Guidelines for Trustworthy AI," Europa.eu. 2019. Acceso mar. 2023. [En línea]. Disponible: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

[2] Unión Europea, "A definition of AI: main capabilities and disciplines," 2019. Acceso mar. 2023. [En línea]. Disponible: https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf

[3] M. Cole, S. Gal, B. Gronlund, S. Hellman, S. Olebo, y H. Prakken, "Politics by Automatic Means? A Critique of Artificial Intelligence Ethics at Work," *Front. Artificial Intell.*, vol. 5, jul. 2022. doi: 10.3389/frai.2022.869114.

[4] P. Moradi y K. Levy, "The Future of Work in the Age of AI: Displacement or Risk-Shifting?", en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 270-288.

[5] B.C. Stahl, "Artificial Intelligence for a Better Future. An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies." Cham: Springer International Publishing, 2021.

[6] P. Boddington, "Normative Modes: Codes and Standards," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. 2020, Oxford, UK: Oxford University Press, p. 124-140.

[7] C. Bartneck, C. Lütge, A. Wagner, and S. Welsh, "Trust and Fairness in AI Systems," en *An Introduction to Ethics in Robotics and AI*, Springer Briefs in Ethics, Cham: Springer, 2021, doi: 10.1007/978-3-030-51110-4_4.

[8] B. Doerr, C. Doerr, y F. Ebel, "From black-box complexity to designing new genetic algorithms," *Theor. Comput. Sci.*, vol. 567, pp. 87-104, 2015. doi: 10.1016/j.tcs.2014.11.028.

[9] J. Mökander, J. Morley, M. Taddeo, y L. Floridi, "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations," *Sci. Eng. Ethics*, vol. 27, no. 4, p. 44, 2021, doi: 10.1007/s11948-021-00319-4.

[10] M. Taddeo, T. McCutcheon, y L. Floridi, "Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed. Cham: Springer International Publishing, 2021, pp. 289-297, doi: 10.1007/978-3-030-81907-1_15.

[11] J. Kroll, "Accountability in Computer Systems," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford, UK: Oxford University Press, 2020, pp. 180-196.

[12] M. Le Bui y S.U. Noble, "We're Missing a Moral Framework of Justice in Artificial Intelligence," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford University Press, 2020, pp. 162-179.

[13] L. Floridi and J. Cowls, "A Unified Framework of Five Principles for AI in Society," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford, UK: Oxford University Press, 2021, pp. 5-17.

[14] C. Bartneck, C. Lütge, A. Wagner, y S. Welsh, "Responsibility and Liability in the Case of AI Systems," in *An Introduction to Ethics in Robotics and AI*, C. Bartneck, et al., Eds. Springer International Publishing, 2021, pp. 39-44.

[15] J. Antoniou y O. Tringides, "Personal Data, Cloud Platforms, Privacy and Quality of Experience," in *Effects of Data Overload on User Quality of Experience*, J. Antoniou and O. Tringides, Eds. Cham: Springer International Publishing, 2023, pp. 37-54.

[16] C. Bartneck, C. Lütge, A. Wagner y S. Welsh, "Privacy Issues of AI," in *An Introduction to Ethics in Robotics and AI*, C. Bartneck, et al., Eds. Cham: Springer International Publishing, 2021, pp. 61-70.

[17] S.H. Vieweg, "Ethical AI Implementation," en *AI for the Good: Artificial Intelligence and Ethics*, S.H. Vieweg, Ed. 2021, Springer International Publishing, p. 227-251.

[18] C. Rushfield y P. Smithsonian Institution Scholarly, "Stemming the Tide: Global Strategies for Sustaining Cultural Heritage through Climate Change. A Smithsonian contribution to knowledge". Washington, D.C: Smithsonian Scholarly Press, 2021.

[19] V. Dignum, "Responsibility and Artificial Intelligence," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford University Press, 2020, pp. 214-231.

[20] Unión Europea, "Policy and investment recommendations for trustworthy Artificial Intelligence," Europa.eu, 2019. Acceso mar. 2023. [En línea]. Disponible: https://bsu.buap.mx/b24

[21] Parlamento Europeo, "Regulación de la inteligencia artificial en la UE: la propuesta del Parlamento," 2022. Acceso mar. 2023. [En línea]. Disponible: https://bsu.buap.mx/b25

[22] Europe Direct, "Presupuesto de la UE para 2023: capacitar a Europa para que siga conformando un mundo cambiante," 2022. Acceso mar. 2023. [En línea]. Disponible: https://bsu.buap.mx/b26

[23] European Commission, “European Education Area,” 2023. Disponible en: https://education.ec.europa.eu/focus-topics

[24] J. Fjeld, N. Achten, H. Hilligoss, A. C. Nagy, y M. Srikumar, “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI,” Berkman Klein Center for Internet & Society, Research Publication No. 2020-1, 2020, doi: 10.2139/ssrn.3518482.

[25] B. Ammanath, “Trustworthy AI: A Business Guide for Navigating Trust and Ethics in AI”. NJ, USA: Wiley, 2022.

[26] J. Meltzer y A. Tielemans, “The European Union AI Act,” 2022. Acceso mar. 2023. [En línea]. Disponible: https://bsu.buap.mx/b3b

[27] S. Fukuda-Parr y E. Gibbons, “Emerging consensus on ‘ethical AI’: human rights critique of stakeholder guidelines,” Global Policy, vol. 12, pp. 32-44, 2021, doi: 10.1111/1758-5899.12965

[28] D. Acemoglu, D. Autor, J. Hazell y P. Restrepo, “Artificial Intelligence and Jobs: Evidence from Online Vacancies,” *J. Labor Econ*., vol. 40, 2022. https://bsu.buap.mx/ciR

[29] A. Ajunwa y R. Schlund, “Algorithms and the Social Organization of Work,” in *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, pp. 804-822, 2020.

[30] Fondo Económico Mundial, “La confianza es la piedra angular de la Ley de Inteligencia Artificial de la UE - De esto se trata,” 2023. Acceso mar. 2023. [En línea]. Disponible: https://bsu.buap.mx/b3c

[31] UE, “Funding for Digital in the 2021-2027 Multiannual Financial Framework,” Acceso mar. 2023. [En línea]. Disponible: https://digital-strategy.ec.europa.eu/en/activities/funding-digital

ASOCIACIÓN EN IA EN BENEFICIO DE LAS PERSONAS Y LA SOCIEDAD, RETOS Y PERSPECTIVAS

Introducción

La PAI es la "Asociación sobre inteligencia artificial en beneficio de las personas y la sociedad" (Partnership on AI to Benefit People and Society) y es una organización sin fines de lucro con sede en San Francisco, California, que reúne a organizaciones académicas, de la sociedad civil, a empresas tecnológicas y de los medios de comunicación para abordar cuestiones sustanciales, básicamente sobre el futuro de la IA, pero también otros importantes retos mundiales como el cambio climático, la alimentación, la desigualdad, la salud y la educación. Tienen cinco programas que contribuyen al desarrollo de recursos, recomendaciones y mejores prácticas para la IA: inteligencia artificial e integridad mediática; IA, trabajo y economía; justicia, transparencia y responsabilidad en aprendizaje automático (ML); investigación y diseño inclusivos; y seguridad para la IA crítica. En este capítulo se destaca que, si bien PAI ha sido criticada por estar dominada en gran medida por grandes empresas tecnológicas, lo que podría limitar su autonomía y autocrítica, es una organización valiosa e influyente en el campo de la ética y la gobernanza de la IA, y desempeña un papel importante en la promoción del desarrollo y el despliegue responsables, no solo de la IA, sino de la tecnología y los medios de comunicación en general.

Los principios éticos que promueve PAI

Partnership on AI [1] es una organización sin fines de lucro fundada en septiembre de 2016 por un grupo de compañías de tecnología líderes, incluyendo Deep Mind, Amazon, Facebook, Google, Microsoft e IBM, con el objetivo de fomentar la colaboración y la investigación en la inteligencia artificial (IA) y explorar cómo la IA puede beneficiar a las personas y a la sociedad en general en un marco ético compartido [2]. En 2023 contaba con 104 socios de 16 países.

Según PAI, su misión es estudiar y formular mejores prácticas en el desarrollo, la implementación y la regulación de la IA, con un enfoque en la ética, la transparencia y la responsabilidad social [2]. La organización reúne a expertos en IA, investigadores, defensores de los derechos civiles y otros interesados en el campo para trabajar juntos en cuestiones críticas relacionadas con el mundo digital. Especialmente se relaciona con las empresas líderes del mercado mundial con quienes trabaja de manera conjunta en sus aplicaciones; entre las principales empresas afiliadas están: Adobe, Amazon, Apple, Deep Mind, Google, IBM, Intel, Intuit, Meta, Microsoft, Samsung y Sony. Medios de comunicación como la BBC de Londres, el New York Times y la Agencia AP; y 60 organizaciones sin fines de lucro, entre las que destaca OpenAI, UNICEF, el Instituto Alan Turing y el instituto Carnegie, entre muchos otros [3].

PAI ha establecido varias iniciativas y grupos de trabajo que se centran en temas como ética y responsabilidad de la IA, inclusión, diversidad, privacidad, seguridad, educación, difusión del conocimiento y colaboración intersectorial. La organización también ha creado un foro para el diálogo público sobre la IA y ha trabajado con gobiernos y organizaciones alrededor del mundo para desarrollar políticas y regulaciones adecuadas para la IA [4].

La asociación publica permanentemente guías como herramientas y recursos para ayudar a los desarrolladores y otros profesionales a implementar prácticas responsables de IA, por lo que ha establecido una serie de grupos de trabajo que se centran en temas específicos tecnológicos.

Los cinco programas que mantienen son: inteligencia artificial e integridad mediática; IA, trabajo y economía; justicia, transparencia y responsabilidad en aprendizaje automático (ML); investigación y diseño inclusivos; y seguridad para la IA crítica. Cada uno de los anteriores ítems con producción específica, artículos y herramientas metodológicas [1].

PAI tiene seis pilares temáticos en los que se funda la asociación y representan conjuntos de temas en los que visualizan los mayores riesgos y oportunidades para la IA:

1. IA de seguridad crítica

Las mejoras en la asistencia sanitaria y el transporte que requieren un alto nivel de seguridad podrían beneficiarse enormemente del uso de la IA. Existe la posibilidad de mejorar la calidad de vida y evitar miles de muertes mediante el uso reflexivo y

estratégico del reconocimiento de patrones, la toma de decisiones automatizada y las tecnologías robóticas [2] [5].

Sin embargo, en los casos en que la IA se utiliza para tomar decisiones junto a los seres humanos o en su lugar, se debe asegurar que sea fiable, segura y respetuosa con los valores y preferencias de las personas a las que afecta.

2. Una IA justa, transparente y responsable

La IA, y especialmente el aprendizaje automático, tiene potencial para añadir valor a la sociedad mediante el reconocimiento de patrones y el análisis de datos, que pueden utilizarse para desarrollar sistemas de diagnóstico y motores de recomendación útiles en campos como la biomedicina, la salud pública, la seguridad, la justicia penal, la educación y la sostenibilidad [2] [6].

Sin embargo, es importante ser conscientes de los posibles sesgos ocultos en los datos utilizados para desarrollar sistemas de IA, así como de otros factores que podrían afectar la calidad de los resultados de los sistemas [7].

3. IA, trabajo y economía

El progreso de la IA tendrá un impacto significativo en el mercado laboral y en la naturaleza del trabajo. Aunque la introducción de nuevas tecnologías promete aportar un valor considerable a la economía, también puede provocar trastornos en el empleo y una reducción de la necesidad de determinados puestos debido a la automatización [2] [8].

Por ello, se está debatiendo cómo minimizar los posibles efectos negativos de la IA en el mercado laboral y garantizar que los beneficios del avance tecnológico se distribuyan de forma equitativa sin dejar de fomentar la competencia y la innovación [9]. Es necesario estudiar y comprender las mejores estrategias para avanzar y participar en los debates en curso [10].

4. Colaboración entre personas y sistemas de IA

La IA puede mejorar la percepción, la cognición y la capacidad de resolución de problemas del ser humano. Algunos ejemplos son las tecnologías de IA que pueden ayudar a los médicos a realizar diagnósticos más precisos y oportunos, y la asistencia de IA proporcionada a los conductores de vehículos para ayudarles a evitar situaciones peligrosas y accidentes [2] [11].

La investigación y el desarrollo en el ámbito de la colaboración entre la IA y el ser humano son necesarios para desarrollar las mejores prácticas. Esto incluye métodos para proporcionar a las personas claridad sobre la comprensión y la confianza que los sistemas de IA tienen sobre aplicaciones específicas, formas de coordinar las contribuciones humanas a la resolución de problemas, y permitir que los sistemas trabajen con las personas de forma armónica.

5. Influencias sociales de la IA

Es un hecho que los avances de la IA afectarán a diversos aspectos de la sociedad y la vida humana, como la privacidad, la democracia, la justicia penal, los medios de comunicación y los derechos humanos. El uso de tecnologías de IA para personalizar la información y ofrecer recomendaciones puede ser beneficioso, pero también existe la posibilidad de que estas tecnologías manipulen involuntaria o intencionadamente a los usuarios e influyan en sus opiniones [2] [12].

Por tanto, es necesario un diálogo abierto y una colaboración reflexiva sobre los posibles efectos de la IA en la sociedad y las personas. El objetivo de PAI es fomentar el debate y la exploración de las formas sutiles y significativas en que la IA podría afectar la cosmovisión de una cultura.

6. IA y bien social

La IA tiene el potencial de promover el bienestar público en áreas como la educación, la vivienda, la salud pública y la sostenibilidad. En Partnership on AI consideran que la colaboración con organizaciones, tanto públicas como privadas, incluidos el mundo académico, las sociedades científicas, las ONGs, los emprendedores sociales y los particulares interesados, puede ayudar a generar debates y catalizar esfuerzos para abordar los retos más acuciantes de la sociedad [2].

Algunos ejemplos de cómo la inteligencia artificial puede ayudar específicamente son: eficiencia energética, detección de fraude eléctrico, vehículos autónomos, gestión del tráfico, salud pública, biología molecular, entre muchos otros campos. De lo anterior, los integrantes de PAI han propuesto seis principios básicos para una ética de la IA:

1. Una IA benéfica

Esta afirmación significa que habrá un esfuerzo para garantizar que las tecnologías de IA se utilicen de forma que tengan un impacto positivo y den poder al mayor

número de personas posible. Sugiere que habrá que centrarse en crear aplicaciones equitativas y accesibles de la IA que beneficien a las personas y a la sociedad en su conjunto, en lugar de beneficiar solo a unos pocos elegidos [2].

2. Escuchar y aprender de todas las voces

PAI hará un esfuerzo por involucrar y comunicarse activamente con el público y las partes interesadas sobre sus iniciativas de IA. La intención de Partnership on AI es educar a la gente sobre el enfoque y los objetivos del trabajo de IA, y buscar opiniones y responder a cualquier pregunta o preocupación que puedan generarse. Esto sugiere un compromiso con la transparencia y la comunicación abierta para garantizar que el desarrollo y la aplicación de las tecnologías de IA se ajuste a las necesidades y los valores de la comunidad en general [2].

3. Implicación de la investigación y diálogo abiertos

Debe existir el compromiso de investigar y facilitar el debate sobre las implicaciones éticas, sociales, económicas y jurídicas de la IA. La intención es garantizar que el desarrollo y el uso de la IA se ajuste a los principios y valores éticos, y abordar de forma proactiva las posibles repercusiones negativas. Esto sugiere un reconocimiento de la necesidad de una toma de decisiones responsable y transparente en relación con el desarrollo y la aplicación de las tecnologías de IA [2].

4. Compromiso activo de todas las partes

Existe la convicción de que la investigación y el desarrollo deben comprometerse activamente y responsabilizarse ante un conjunto diverso de partes interesadas. Estas partes incluyen a individuos, organizaciones y comunidades que pueden verse afectados por la creación y utilización de tecnologías de IA. El objetivo es garantizar que la IA responda a las necesidades e inquietudes de un amplio espectro de personas y grupos, y que las ventajas se repartan equitativamente entre toda la sociedad. [2].

5. Compromiso y atención a todas las partes involucradas

Existe el compromiso de colaborar con la comunidad empresarial en el desarrollo y la aplicación de las tecnologías de IA. La intención es garantizar que se comprendan y aborden las preocupaciones y oportunidades específicas de las distintas industrias y sectores [2].

La inclusión de representantes de la comunidad empresarial ofrece la oportunidad de trabajar en colaboración para identificar posibles casos de uso y áreas de innovación. Esto sugiere un reconocimiento de la importancia de implicar a todas las partes interesadas en el desarrollo y la aplicación de las tecnologías de IA.

6. Maximizar beneficios y enfrentar los retos

Trabajando para preservar la privacidad y la seguridad de las personas, PAI se esfuerza por aprovechar las ventajas de la tecnología de IA y gestionar las posibles dificultades, intentando comprender y respetar los intereses de todas las partes a las que puedan afectar los posibles avances de la IA [2]. PAI trabaja para mantener el compromiso social de los grupos de investigación e ingeniería de IA sobre las posibles repercusiones de la tecnología en general. Garantizar que la investigación y la tecnología de la IA sean sólidas, dignas de confianza y fiables, y que funcionen dentro de límites seguros. Oponerse a la creación y aplicación de tecnología de IA que pueda violar los derechos humanos o los acuerdos internacionales, apoyando al mismo tiempo las medidas de seguridad y las tecnologías beneficiosas. Precisamente del seno de la PAI surgió en marzo de 2023 la propuesta de hacer una tregua y frenar el desarrollo de la IA, más allá de GPT4, mientras se desconozcan sus implicaciones y riesgos a largo plazo [21].

7. Inteligibilidad y explicabilidad

Es fundamental que las personas comprendan e interpreten el funcionamiento de los sistemas de IA para explicar eficazmente su tecnología. Comprender los sistemas de IA es un primer paso hacia la mejora de la accesibilidad y el fomento de la confianza entre las personas que interactúan abiertamente con ellos. Ofrecer explicaciones no solo promueve la aceptación pública, sino que también ayuda a identificar y resolver posibles sesgos o errores [2].

8. Cooperación, confianza y apertura entre científicos

Fomentar una cultura de colaboración y transparencia entre los científicos e ingenieros de IA para alcanzar sus objetivos compartidos [2]. La colaboración y la confianza entre científicos e ingenieros de IA son esenciales para avanzar en el desarrollo de la tecnología de forma ética y sostenible. Una cultura abierta promueve el intercambio de conocimientos, la crítica constructiva y la revisión por pares, lo que puede conducir a mejores soluciones y minimizar los riesgos de consecuencias no deseadas. En última instancia, estos esfuerzos también pueden contribuir a fomentar la confianza pública en la IA y promover su uso responsable.

Haciendo que la IA sea inclusiva

Además de lo anterior, en el documento "Haciendo que la IA sea inclusiva, cuatro principios rectores para el compromiso ético" [13] de Tina M. Park de julio de 2022, PAI ha propuesto:

1. Trabajo participativo

Reconocer el valor de la participación de los usuarios y del público es crucial para fomentar un enfoque integrador del trabajo. Es imperativo que todo el mundo tenga la oportunidad de disfrutar de los beneficios que aporta la IA [13].

Para construir sistemas de IA se necesitan grandes cantidades de datos, por lo que la tecnología depende mucho de la participación del público. Los datos que los usuarios proporcionan son necesarios para que la tecnología pueda funcionar correctamente y generar beneficios [14].

La participación se refiere a cualquier contribución directa o indirecta a la creación, desarrollo, implementación y sostenibilidad de un sistema de IA. Esto significa que cualquier acción que ayude a crear o mejorar los sistemas se considera una forma de participación. También lo son la provisión de datos, la retroalimentación sobre el uso de los sistemas o la capacitación para su desarrollo [15]. La producción de textos, fotografías, videos, audios, y etiquetado entran en la misma categoría.

La importancia de la participación pública en la creación de estos sistemas radica en que están diseñados para satisfacer las necesidades y expectativas de los usuarios. Sin la participación de una amplia gama de personas que representen diversas perspectivas y necesidades, no se puede lograr esto [16].

2. Terminar con asimetrías

Para Park [1] la participación de las partes interesadas debe abordar las asimetrías de poder inherentes. Muchos ciudadanos dudan a la hora de apoyar iniciativas participativas dirigidas por empresas u otras entidades, especialmente porque hay escepticismo en las comunidades históricamente oprimidas y explotadas [13].

En el pasado, los grupos como la comunidad afroamericana de Estados Unidos han sido tratados injustamente, lo que ha llevado a que tengan desconfianza hacia los sistemas tecnológicos que se desarrollan. Por esta razón, pedirles que participen

en la creación de sistemas de IA puede generar suspicacia, ya que sienten que se les está pidiendo que trabajen sin remuneración para beneficiar a otros, en lugar de ser incluidos como verdaderos socios en el desarrollo de la tecnología [1]. Cosa que, dicho sea, es verdad.

Los usuarios y el público en general carecen a menudo de la capacidad de obtener información privilegiada o de influir significativamente en las decisiones cruciales que toman las organizaciones de IA y aprendizaje automático. Los desarrolladores, que actúan como intermediarios, tienen el poder de controlar el acceso a la información y tomar decisiones importantes. Sin embargo, incluso los propios desarrolladores pueden no tener la autoridad última para conceder dicho acceso. Además, los desequilibrios de poder derivados de factores históricos como la segregación y la discriminación contribuyen a estas relaciones inequitativas [1].

Reconocer y abordar esta dinámica es esencial para establecer relaciones respetuosas y mutuamente beneficiosas entre quienes promueven las iniciativas de IA y los miembros de la comunidad implicados. Es crucial reconocer el impacto de factores sociales como el racismo y la misoginia, ya que estas dinámicas pueden tener un efecto continuo en la relación, aunque las interacciones interpersonales parezcan más equitativas [1].

3. Inclusión y participación

La inclusión y la participación pueden integrarse en todas las fases del ciclo de vida del sistema. Sigue siendo frecuente encontrar prácticas participativas implementadas en el desarrollo de IA hacia el final del ciclo de vida, en lugar de estar integradas a lo largo de todo el proceso. En otras palabras, estas prácticas participativas son, a menudo, una ocurrencia tardía en lugar de un componente central del desarrollo de la IA [13].

Debido al enfoque anterior es esencial garantizar que las prácticas participativas se integren plenamente en todo el ciclo de vida del desarrollo de la IA para promover un proceso de desarrollo más inclusivo y colaborativo que tenga en cuenta las necesidades y perspectivas de todas las partes interesadas, no solo cuando el producto ya está hecho [16].

Se afirma que las personas que abogan por la equidad y la inclusión han destacado la importancia de implicar a los usuarios y a los miembros de las comunidades afectadas en todas las fases del proceso de desarrollo de los proyectos de IA. Esto significa que las aportaciones y comentarios de todas las partes interesadas deben

buscarse e incorporarse desde las primeras fases del proyecto hasta el despliegue final.

La manera más eficaz de lograrlo es aplicando prácticas participativas integradoras que creen relaciones significativas entre todas las partes. Al darles el espacio para dirigir el propósito y la intención del proyecto de IA los miembros de la comunidad pueden estar facultados para garantizar que el proyecto se alinea con sus valores y prioridades, lo que puede dar lugar a resultados más éticos y equitativos [18].

4. Transversalidad inclusiva

Según Park [1], incorporar la inclusión y la participación es de vital importancia a la hora de aplicar otros principios de la IA responsable. A menudo, los debates sobre el uso ético de la IA se centran en normas individuales por separado, pasando por alto la interconexión de estos principios [13]. Por ejemplo, una norma puede hacer hincapié en el acceso universal a la tecnología de la IA, mientras que otra destaca la importancia de un uso justo. Si bien estas normas tienen importancia por sí mismas, es igualmente vital que estén interconectadas y se apoyen mutuamente, trabajando en armonía para lograr prácticas de IA responsables.

Conclusiones parciales

De acuerdo con lo planteado por Partnership on AI para maximizar los beneficios y afrontar los retos de la IA, es crucial garantizar que los sistemas sean favorables para toda la sociedad. Esto requiere el compromiso activo de todas las partes involucradas en su desarrollo y despliegue, incluidos científicos, ingenieros, distribuidores, políticos y usuarios. Alcanzar estos objetivos también requiere una investigación que tenga en cuenta las implicaciones de la tecnología de IA para la sociedad y el medio ambiente.

Para lograr una comprensión holística de las consecuencias de la IA, es esencial escuchar y aprender de todas las voces del sector, incluidas las de procedencias y perspectivas que no estén a favor del desarrollo de la tecnología. Esto significa crear una cultura de apertura y cooperación entre los científicos e ingenieros de IA, así como colaborar con comunidades más amplias y expertos de otras disciplinas, por ejemplo, ética, filosofía, derecho, sociología, antropología, economía y ciencias cognitivas.

PAI ha sido criticada porque, aunque la organización incluye a una amplia gama de sectores representados, sigue estando dominada en gran medida por las grandes empresas tecnológicas, lo que puede limitar su capacidad para ser plenamente autónoma, autocrítica y representar los intereses globales [19].

También se ha señalada de ser una panacea para protegerse las grandes empresas de un verdadero marco regulatorio internacional [22]. Otra crítica es que las iniciativas y proyectos de PAI tienden a centrarse en la autorregulación voluntaria y las normas impulsadas por la industria, en lugar de reglamentos jurídicamente vinculantes y supervisión gubernamental. Sin una normativa y supervisión más estrictas, las empresas pueden no tener que rendir cuentas de sus actos y dar prioridad a los beneficios económicos frente a las consideraciones éticas [20].

A pesar de estas críticas, PAI sigue siendo una organización influyente y necesaria en el campo de la ética y la gobernanza de la IA, y desempeña un papel decisivo en la promoción del desarrollo y el despliegue responsables de la IA.

En conclusión, PAI representa una iniciativa prometedora que puede ayudar a afrontar los retos y liberar el vasto potencial de la IA para la mejora de la humanidad. Al fomentar la colaboración y la cooperación entre las partes interesadas de diversos sectores y regiones, la asociación puede contribuir al desarrollo de una IA responsable, ética e integradora. Gracias a la dedicación persistente y al compromiso colectivo de todos los implicados, el futuro de la IA, desde PAI, es prometedor para servir al bien común y mejorar el bienestar humano. Esta organización sin ánimo de lucro, impulsada por su misión, sigue dedicada a fomentar la colaboración entre diversas voces de distintos sectores, disciplinas y grupos demográficos para cumplir con su objetivo primordial que es que los avances en IA produzcan resultados positivos, tanto para las personas como para la sociedad en su conjunto [2].

Referencias

[1] Partnership on AI, “Partnership on AI is bringing together diverse voice from across the AI community,” Partnershiponai.org. [En línea]. Acceso mar. 2023. [En línea]. Disponible: https://partnershiponai.org/

[2] Partnership on AI, “About us,” Partnershiponai.org. Acceso mar. 2023. [En línea]. Disponible: https://partnershiponai.org/about/

[3] Partnership on AI, “Our Funding. PAI relies on a multitude of funding sources to accomplish our goals,” Partnershiponai.org. Acceso mar. 2023. [En línea]. Disponible: https://partnershiponai.org/funding/

[4] Partnership on AI, “Our Resource Library. Our collected papers, resources, and other outputs,” Partnershiponai.org. Acceso mar. 2023. [En línea]. Disponible: Disponible en: https://partnershiponai.org/resources/

[5] Partnership on AI, “Safety Critical AI,” Partnershiponai.org. Acceso mar. 2023. [En línea]. Disponible: https://partnershiponai.org/program/safety-critical-ai/

[6] Partnership on AI, “Fairness, Transparency, and Accountability & About ML,” Partnershiponai.org. Acceso mar. 2023. [En línea]. Disponible: https://partnershiponai.org/program/fairness-transparency-and-accountability-about-ml/

[7] J. Fjeld, N. Achten, H. Hilligoss, A. C. Nagy, y M. Srikumar, “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI,” Berkman Klein Center for Internet & Society, Research Publication No. 2020-1, 2020, doi: 10.2139/ssrn.3518482

[8] Partnership on AI, “AI, Labor, and the Economy,” Partnershiponai.org. Acceso mar. 2023. [En línea]. Disponible: Disponible en: https://partnershiponai.org/program/ai-labor-and-the-economy/

[9] D. Acemoglu, D. Autor, J. Hazell y P. Restrepo, “Artificial Intelligence and Jobs: Evidence from Online Vacancies,” *J. Labor Econ.*, vol. 40, 2022. Disponible: https://bsu.buap.mx/ciR

[10] P. Moradi y K. Levy, “The Future of Work in the Age of AI: Displacement or Risk-Shifting?,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 270-288.

[11] Partnership on AI, “PAI Research Promotes Responsible Collaborations between People and AI Systems,” Partnershiponai.org. Acceso mar. 2023. [En línea]. Disponible: Disponible en: https://partnershiponai.org/cpais-research/

[12] Partnership on AI, “Fairer Algorithmic Decision-Making & Its Consequences,” Partnershiponai.org. Acceso mar. 2023. [En línea]. Disponible: https://partnershiponai.org/paper/fairer-algorithmic-decision-making-and-its-consequences/.

[13] T. Park, “Making AI Inclusive: 4 Guiding Principles for Ethical Engagement,” Partnershiponai.org. Acceso mar. 2023. [En línea]. Disponible: https://partnershiponai.org/paper/making-ai-inclusive-4-guiding-principles-for-ethical-engagement/.

[14] L. Floridi, “Establishing the Rules for Building Trustworthy AI,” en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Springer Intern. Publish., 2021, pp. 41-45.

[15] L. Floridi et al., “An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations,” en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Springer Intern. Publish., 2021, pp. 19-39, doi: 10.1007/s11023-018-9482-5.

[16] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo y L. Floridi, “The ethics of algorithms: key problems and solutions,” *AI & Soc.*, vol. 37, no. 1, pp. 215-230, Feb. 2022. doi: 10.1007/s00146-021-01154-8.

[17] M. Taddeo, T. McCutcheon, y L. Floridi, "Trusting artificial intelligence in cybersecurity is a double-edged sword," *Nat. Mach. Intell.*, vol. 1, pp. 557-560, 2019.

[18] S. Milano, M. Taddeo, y L. Floridi, "Recommender systems and their ethical challenges," *AI & SOC.*, vol. 35, no. 4, pp. 957-967, 2020, doi: 10.1007/s00146-020-00950-y.

[19] R. Ochigame, "The invention of 'ethical AI'," Acceso mar. 2023. [En línea]. Disponible: https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/

[20] Rességuier y R. Rodrigues, "AI ethics should not remain toothless! A call to bring back the teeth of ethics," *Big Data & Soc.*, vol. 7, no. 2, p. 205395172094254l, 2020.

[21] Future of Life Institute, "Pause Giant AI Experiments: An Open Letter," Futureoflife.org. Acceso jun. 2023. [En línea]. Disponible: https://bsu.buap.mx/b8d

[22] T. Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Mind and Mach.*, vol. 30, no. 1, pp. 99-120, 2020. DOI: 10.1007/s11023-020-09517-8

IEEE: UN ESTÁNDAR GLOBAL COMO INICIATIVA ÉTICA DE LA IA

Introducción

"Diseño alineado éticamente: una visión para priorizar el bienestar humano con sistemas autónomos e inteligentes, primera edición" fue redactado bajo la "Iniciativa global sobre ética de los sistemas autónomos e inteligentes" y es un documento que pretende ofrecer un marco para las consideraciones éticas en el diseño, desarrollo, despliegue y uso de los sistemas de IA, con el objetivo de garantizar que las tecnologías digitales ayuden a las personas en sus labores. El Diseño consta de ocho principios: derechos humanos, bienestar, agencia de datos, eficacia, transparencia, rendición de cuentas, conciencia de uso indebido y competencia. Algunas de las observaciones a la iniciativa son que podría estar influida por los beneficios de la industria, y no represente las perspectivas e intereses de las comunidades digitalmente marginadas. A pesar de estas críticas, el documento del IEEE sigue siendo influyente en el campo de la ética y la gobernanza de la IA, y ha desempeñado un papel importante en la configuración del debate mundial en torno al desarrollo y despliegue responsables de la tecnología. También se analiza que el Instituto ha lanzado el "IEEE 7000™-2021, proceso de modelo estándar IEEE para abordar preocupaciones éticas durante el diseño del sistema" con el que sienta el precedente más objetivo de una ética de la IA aplicable.

La iniciativa global de IEEE

"Diseño alineado éticamente: una visión para priorizar el bienestar humano con sistemas autónomos e inteligentes, primera edición" (Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition) redactado bajo la "Iniciativa global sobre ética de los sistemas autónomos e inteligentes" (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems) [1] ha influido en el desarrollo de la ética y la gobernanza de los sistemas de IA, y ha sido adoptada por varias organizaciones y gobiernos del mundo como guía para el despliegue responsables de la IA.

El Instituto de ingenieros eléctricos y electrónicos (IEEE) es la mayor organización profesional técnica dedicada al avance de la tecnología en beneficio de la humanidad. Sus fundadores fueron, entre otros, Thomas Alva Edison y Alexander Graham Bell [2].

El IEEE cuenta con más de 423 000 miembros en más de 160 países, según su sitio web [3]. Entre sus miembros hay ingenieros, científicos y otros profesionales de la tecnología, así como estudiantes y educadores. La organización está conformada por varias sociedades, cada una de las cuales se centra en un área técnica específica, como la electricidad, la energía, las telecomunicaciones y, especialmente, la informática.

El IEEE puso en marcha la "Iniciativa global para la ética de los sistemas autónomos e inteligentes" (Global Initiative on Ethics of Autonomous and Intelligent Systems) [1] con el fin de abordar los problemas éticos relacionados con la creación y difusión de dichos sistemas. En la iniciativa se identificaron más de 120 cuestiones significativas y se sugirieron posibles soluciones [4]. Además, ha servido de base para 14 proyectos normalizados que están actualmente en curso a través de la Asociación de normas del IEEE. Los principios en el diseño éticamente alineado son:

1. Derechos humanos

> Los sistemas de IA deben ser creados y operados para respetar, promover y proteger los derechos humanos reconocidos internacionalmente *[1, p. 19].*

Para garantizar que el uso de sistemas de IA no vulnera los derechos humanos, las libertades, la dignidad o la privacidad, deben establecerse marcos de gobernanza, incluidas normas y organizaciones reguladoras. Estos procesos también deben proporcionar trazabilidad. Esto contribuirá a aumentar la confianza del público en la IA [1].

Es necesario encontrar un mecanismo para convertir las consideraciones políticas y tecnológicas existentes y futuras, en obligaciones legales apegadas a derecho. Un procedimiento de este tipo debe tener en cuenta diversas normas culturales, así como diversos marcos jurídicos y normativos [5].

2. Bienestar

> Los creadores de sistemas autónomos e inteligentes adoptarán el aumento del bienestar humano como principal criterio de éxito para el desarrollo *[1, p. 21].*

Las mejores y más utilizadas mediciones de bienestar deberían aplicarse como referencia para los sistemas de IA, con el fin de garantizar que sea la prioridad como resultado en todos los diseños [6]. La declaración sugiere que, a la hora de diseñar e implantar la IA y los sistemas autónomos, es importante dar prioridad al bienestar humano como objetivo último. Para lograrlo, deben aplicarse indicadores ampliamente aceptados como referencia para evaluar el impacto de estos sistemas. Si bien esto no resulta tan sencillo en muchos casos, se puede crear las métricas adecuadas que contabilicen, de alguna manera, los impactos positivos y negativos que la tecnología pueda tener en los usuarios. Si se deja al azar, los aspectos perjudiciales podrían ser detectados cuando ya se es demasiado tarde.

3. Control de los datos

> Los creadores de sistemas autónomos e inteligentes deberán dotar a las personas de la capacidad de acceder a sus datos y compartirlos de forma segura, para mantener la facultad de las personas de tener control sobre su identidad *[1, p.23].*

Para IEEE los gobiernos y otras organizaciones deberían emprender esfuerzos para investigar, probar y aplicar tecnologías y procedimientos que permitan a los usuarios tener control sobre sus datos personales, en concreto permitiéndoles decidir caso por caso quién puede acceder a sus datos y procesarlos y con qué fines específicos [7]. Es crucial explorar si las estrategias de tutela existentes para niños y personas con capacidad de decisión disminuida se ajustan a esta recomendación o si alguien más debiese asumir la responsabilidad [8].

El ser humano debe controlar a los sistemas de IA sobre el modo en que toman decisiones, aprenden y el manejo que hagan de la información personal ante terceros. Por ejemplo, en todo momento permitir a los usuarios eliminar sus datos y su cuenta por completo.

4. Eficacia

> Los creadores y operadores deberán aportar pruebas de la eficacia y adecuación a los fines de los sistemas autónomos e inteligentes *[1, p. 25].*

Los responsables del diseño y la implantación de los sistemas de IA deben demostrar mediante pruebas tangibles que los sistemas son capaces de alcanzar los objetivos previstos y son adecuados para el fin perseguido. De acuerdo con la IEEE, la creación de sistemas de IA debe tener como objetivo la identificación de métricas o puntos de referencia que sirvan como indicadores fiables del éxito del sistema en la consecución de sus objetivos, el cumplimiento de las normas y el funcionamiento dentro de las tolerancias de riesgo [9]. Los diseñadores de sistemas de IA deben asegurarse de que todas las partes interesadas, como usuarios, certificadores de seguridad y reguladores del sistema, puedan acceder fácilmente a los resultados cuando se apliquen las métricas establecidas [10].

Considérense los sistemas de IA para la conducción autónoma, los desarrolladores y operadores deben aportar pruebas que demuestren que el software es eficaz para percibir con precisión el entorno, tomar decisiones oportunas y garantizar la seguridad de pasajeros y peatones. Estas pruebas deben incluir resultados de pretests rigurosos, estudios de simulación, datos de rendimiento en el mundo real y el cumplimiento de las normas de seguridad pertinentes antes de salir al mercado. Si bien hay lugares en que las normativas gubernamentales son más laxas, se debe tener altos estándares éticos en cada proyecto.

5. Transparencia

> La base de una decisión concreta de un sistema autónomo e inteligente debe poder descubrirse siempre *[1, p. 27].*

La IEEE considera que, al garantizar la detectabilidad, los desarrolladores y operadores de los sistemas de IA promueven la responsabilidad, la equidad y la capacidad de identificar y rectificar cualquier sesgo, error o consecuencia no deseada. La transparencia permite evaluar el proceso de toma de decisiones, lo que es crucial para la creación de confianza, la auditoría y la trazabilidad general.

Por ejemplo, los usuarios de robots domésticos o asistenciales deberían disponer de un botón que, al oprimirlo, el robot explique la acción que acaba de realizar [11]. También deben contar con un almacenamiento seguro de los datos de los sensores y del estado interno, similar al de una grabadora de datos de vuelo [12]. El Internet de las cosas ha levantado suspicacias precisamente por la opacidad en algunos casos, por ejemplo, en que se recojan y compartan datos personales sin el consentimiento explícito del usuario.

6. Responsabilidad

> Los sistemas autónomos e inteligentes deben crearse y gestionarse de forma que ofrezcan una justificación inequívoca de las decisiones adoptadas *[1, p. 29].*

Los sistemas autónomos e inteligentes pueden emprender acciones y tomar decisiones sin supervisión humana directa. Esto incluye los coches que se conducen solos, la robótica, los sistemas comerciales automatizados, etcétera. Estos sistemas deben crearse y gestionarse de forma responsable. Su diseño, desarrollo y despliegue deben seguir principios éticos capaces de justificar y explicar sus decisiones y acciones [5]. Las explicaciones deben ser claras e inequívocas, no vagas, contradictorias ni susceptibles de múltiples interpretaciones. No debe haber dudas sobre la lógica del comportamiento del sistema.

Esta explicabilidad es importante para mantener la responsabilidad, depurar el sistema, evitar sesgos involuntarios, generar confianza entre los usuarios y mucho más. Si un sistema no puede explicarse con claridad, resulta difícil evaluar si funciona de forma segura, ética o según lo previsto [6].

Esto es especialmente importante porque estas tecnologías son relativamente nuevas y aún no se conocen las repercusiones a futuro [13].

7. Conciencia del mal uso

> Los creadores deberán protegerse contra todos los usos indebidos y riesgos potenciales de sus sistemas autónomos e inteligentes en funcionamiento *[1, p. 31].*

Los diseñadores de sistemas de IA deben conocer las técnicas habituales de abuso y tratar de no hacerlas accesibles en sus diseños [14]. Ofrecer instrucción ética y hacer conciencia de los posibles riesgos del uso indebido de la tecnología de IA, y sensibilizar al público sobre las repercusiones. Por ejemplo, la creación de imágenes ficticias puede ser muy cómico para algunos, pero puede traer daños graves a la sociedad, a las personas involucradas y a la economía, aunque, en realidad, todavía no se pueda saber que alcances habrá de tener este tipo de fenómenos [7].

8. Competencia

> Los creadores harán las especificaciones y los operadores deberán respetar los conocimientos y destreza necesarios para un funcionamiento seguro y eficaz *[1, p. 32].*

Los tipos y grados de conocimiento necesarios para comprender y utilizar cualquier aplicación de IA deben ser especificados por sus creadores. Deben identificar los conocimientos requeridos, tanto para el sistema en su conjunto como para cada componente individual [15]. Los operadores de IA deben establecer políticas escritas que especifiquen cómo debe utilizarse. Estas directrices deben cubrir los usos prácticos de los sistemas, los requisitos previos para su uso eficaz, quién es competente para manejarlos, qué formación es necesaria para los operadores, cómo evaluar rendimiento del sistema y qué resultados deben esperarse. Las políticas también deben especificar las situaciones en las que puede ser necesario que el operador anule o detenga al sistema mismo [16].

Las recomendaciones anteriores están encaminadas a crear un estándar y conforman la primera serie de normas para una directriz tipo ISO sobre ética en la IA denominada serie IEEE P7000™ de proyectos de estandarización.

Norma IEEE 7000™ 2021. Preocupaciones éticas durante el diseño de sistemas

Más de 150 expertos trabajaron en la "Norma 7000™-2021 Proceso del modelo estándar IEEE para abordar las preocupaciones éticas durante el diseño del sistema" (7000™-2021IEEE Standard Model Process for Addressing Ethical Concerns during System Design) [17] y fue el resultado de debates en línea que tuvieron lugar a lo largo de cinco años y en los que participaron representantes de Europa, Oriente Medio, Estados Unidos, Australia y América Latina, así como especialistas de distintas disciplinas entre las que destacan ingenieros en sistemas, filósofos, consultores y abogados, entre otras áreas.

El modelo de proceso estándar del IEEE tiene como objetivo combinar temas éticos con la práctica para reducir riesgos e impulsar la innovación de la ingeniería de sistemas dentro de un enfoque compartido.

Para la IEEE ignorar los valores del usuario es un peligro en el diseño de ingeniería. Muchos bienes y servicios funcionan con sistemas de IA, que son algoritmos que operan "por debajo" del sistema y que tienen un impacto significativo en los datos, las identidades y los valores de los usuarios [18]. A pesar de los mejores esfuerzos de un fabricante, un proceso de diseño enfatizará las convicciones de sus creadores y su solvencia ética no siempre estará garantizada. En la era de los algoritmos, la innovación responsable requiere un enfoque basado en principios que vayan más allá de la ingeniería de sistemas convencional [19].

El estándar IEEE 7000™-2021 ofrece precisamente estos valores a las empresas a través de un método práctico para superar los problemas asociados a su transformación digital. La metodología ofrece una perspectiva más amplia para tener en cuenta los posibles daños causados por el diseño de productos o sistemas que no estén bien calibrados [20].

En el contexto del aprendizaje automático, se dice que un algoritmo está mal calibrado si sus predicciones o estimaciones son sistemáticamente sesgadas o inexactas. Más concretamente, un algoritmo mal calibrado puede producir estimaciones o predicciones que son, en promedio, demasiado altas o bajas en comparación con los valores reales [21]. Por ejemplo, considérese un escenario en el que un algoritmo está entrenado para evaluar la probabilidad de un resultado binario, como si un paciente tiene o no una enfermedad específica. Si el algoritmo sobrestima

sistemáticamente esta posibilidad, puede indicar erróneamente que está enfermo cuando, en realidad, no lo está [22].

Conclusiones parciales

La iniciativa de IEEE proporciona directrices éticas que se aplican a todo tipo de sistemas autónomos e inteligentes, incluidos los robots mecánicos, así como los robots algorítmicos, autos de conducción autónoma, sistemas de software, sistemas de diagnóstico médico, asistentes personales inteligentes y bots algorítmicos de chat, en diferentes entornos, tanto reales como virtuales, contextuales y de realidad mixta en donde la IA esté presente.

El texto de la Iniciativa y el Estándar 7000™-2021 pretenden ofrecer un marco para las consideraciones éticas específicas en el diseño, desarrollo, despliegue y uso de los sistemas de IA, con el objetivo de garantizar que estas tecnologías beneficien a la humanidad.

Algunas críticas al documento del IEEE es que, nuevamente como en otras iniciativas, los criterios son demasiado amplios y carecen de orientaciones específicas sobre cómo aplicar en la práctica los principios que esboza. Esto puede dificultar que organizaciones y gobiernos traduzcan los principios en acciones y políticas concretas. Sin embargo, también debe decirse que la IEEE es la organización que ha sido más incisiva al crear las normas correspondientes al desarrollo tecnológico, y prueba de ello es la “Norma 7000TM-2021 Proceso del modelo estándar IEEE para abordar las preocupaciones éticas durante el diseño del sistema” que no se debe subestimar como un esfuerzo puntual de implementación de la ética en la IA.

Por supuesto, también se ha expresado la preocupación de que los documentos del IEEE puedan estar influidos por los intereses de las partes interesadas de la industria, como las empresas tecnológicas y no representen plenamente las perspectivas e intereses de todas las partes, en particular las de las comunidades digitalmente marginadas.

A pesar de estas críticas, el Estándar 7000™-2021 y la Iniciativa son influyentes en el campo de la ética y la gobernanza de la IA. La IEEE desempeña un papel sustancial en la configuración del debate mundial en torno al desarrollo y despliegue responsables, no solo de la IA, sino del progreso científico de la humanidad.

Referencias

[23] IEEE. “Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition.” IEEE.org. Acceso abr. 2023. [En línea]. Disponible: https://bsu.buap.mx/b3f

[24] IEEE. “History of IEEE.” IEEE.org. Acceso abr. 2023. [En línea]. Disponible: https://www.ieee.org/about/ieee-history.html.

[25] IEEE. “Membership.” IEEE.org. Acceso abr. 2023. [En línea]. Disponible: https://www.ieee.org/membership/index.html.

[26] R. Chatila y J.C. Havens, “The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems,” en Robotics and Well-Being, M.I. Aldinhas Ferreira et al., Eds., Springer International Publishing, 2019, pp. 11-16, doi: 10.1007/978-3-030-12524-0_2

[27] T. Tzimas, “Legal and Ethical Challenges of Artificial Intelligence from an International Law Perspective.” Cham: Springer, 2021.

[28] M. Dubber, F. Pasquale, y S. Das, Eds. *The Oxford Handbook of Ethics of AI*. Oxford, UK: Oxford University Press, 2020.

[29] J. Antoniou y O. Tringides, “Personal Data, Cloud Platforms, Privacy and Quality of Experience,” en *Effects of Data Overload on User Quality of Experience*, J. Antoniou and O. Tringides, Eds., Cham: Springer International Publishing, 2023, pp. 37-54, doi: 10.1007/978-3-031-06870-6_3.

[30] S. Milano, M. Taddeo, y L. Floridi, “Recommender systems and their ethical challenges,” *AI & Soc.*, vol. 35, no. 4, pp. 957-967, 2020, doi: 10.1007/s00146-020-00950-y.

[31] T. Winkle, “Product Development within Artificial Intelligence, Ethics and Legal Risk.” Alemania: Springer Vieweg, 2022.

[32] L. Floridi, “Ethics, Governance, and Policies in Artificial Intelligence.” Cham: Springer, 2021.

[33] M. Coeckelbergh, “Robot ethics.” MA, USA: MIT Press, 2022.

[34] T. Hauer, “Incompleteness of moral choice and evolution towards fully autonomous AI,” Humanit. and soc. sciences commun., vol. 9, no. 1, p. 38, 2022. Disponible: https://bsu.buap.mx/ciT

[35] J. Bryson, “The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 2-25.

[36] K. Kumari y J. P. Singh, “AI ML NIT Patna at HASOC 2019: Deep Learning Approach for Identification of Abusive Content,” en *Proceedings of Forum for Information Retrieval Evaluation*, vol. 2517, pp. 328-335, 2019. Disponible: https://ceur-ws.org/Vol-2517/T3-20.pdf

[37] M. Groß, “Yes, AI Can: The Artificial Intelligence Gold Rush Between Optimistic HR Software Providers, Skeptical HR Managers, and Corporate Ethical Virtues,” en *AI for the Good: Artificial Intelligence and Ethics*, S.H. Vieweg, Ed., Cham: Springer International Publishing, 2021, pp. 191-225.

[38] B. Zhang et al., “Ethics and Governance of Artificial Intelligence: A Survey of Machine Learning Researchers,” en *IJCAI International Joint Conference on Artificial Intelligence*, 2022.

[39] IEEE, “7000-2021 - IEEE Standard Model Process for Addressing Ethical Concerns during System Design,” IEEE.org. Acceso abr. 2023. [En línea]. Disponible: https://ieeexplore.ieee.org/document/9536679.

[40] L. Floridi, “Establishing the Rules for Building Trustworthy AI,” en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer, 2021, pp. 41-45.

[41] L. Floridi y J. Cowls, “A Unified Framework of Five Principles for AI in Society,” en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer International Publishing, 2021, pp. 5-17.

[42] B.C. Stahl, “AI Ecosystems for Human Flourishing: The Recommendations,” en *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*, B.C. Stahl, Ed., Springer International Publishing, 2021, pp. 91-115.

[43] S. Russell y P. Norvig, “Philosophy, ethics, and safety of AI,” en *Artificial Intelligence: A Modern Approach*, Londres: Pearson, 2022, pp. 1032-1062.

[44] G.Z. Yang, et al., “The grand challenges of Science Robotics,” Sci Robot., vol. 3, no. 14, p. eaar7650, ene. 2018, doi: 10.1126/scirobotics.aar7650. PMID: 33141701.

ÉTICA DE LA IA DESDE LAS EMPRESAS GLOBALES: MICROSOFT, GOOGLE, META Y APPLE

Introducción

En este capítulo se analizan las propuestas éticas para el desarrollo digital y empresarial de cuatro grandes corporativos internacionales: Microsoft, Google (Alphabet), Facebook (Meta) y Apple. Se ponderan cada uno de sus compromisos publicados en sus plataformas respectivas o las políticas compartidas por sus direcciones ejecutivas. Si bien cada una de las megaempresas, al menos en el papel, presume una serie de valores incuestionables por su integridad, también es cierto que la mayoría ha tenido que enfrentar crisis por la carencia precisamente de algunos de los principios por ellos mismos proclamados. También es cierto que no puede hacerse generalizaciones precipitadas, por eso aquí se comparte primero, objetivamente, lo que proponen y, siendo autorregulatorias, se discute hasta qué punto han cumplido sus propias aspiraciones éticas.

El poder fáctico en el mundo

Las grandes empresas tecnológicas como Microsoft, Google, Facebook y Apple han pasado a lo largo de su vida corporativa por diversas etapas y retos frente a la adaptación de sus políticas comerciales a las cambiantes reglas del juego; sin embargo, en el camino, algunas de ellas han tenido que afrontar demandas legales y de la opinión pública que han llegado a comprometer no solo su reputación, sino su valor en el mercado.

Algunas empresas, mejor que otras, han sabido cómo continuar, solventar las observaciones e incluso salir fortalecidas, otras siguen aprendiendo de los golpes y ajustando sus políticas a un entorno social cada vez más incisivo e informado. El público es más perceptivo y muchas veces logra distinguir la intencionalidad de las operaciones de las empresas.

El panorama que rodea el desarrollo de la inteligencia artificial está experimentando cambios significativos. Dada esta evolución, resulta imperativo articular explícitamente los principios éticos sobre los que se construyen los valores fundamentales de estas cuatro grandes empresas. Al esbozar claramente estos pilares axiológicos subyacentes, podemos establecer una base que alinee los avances de la IA con consideraciones éticas, garantizando prácticas responsables y que rindan cuentas en este ámbito de incertidumbre.

Microsoft

La misión de Microsoft es, en sus propias palabras, “empoderar a otros para alcanzar sus metas” [1, p. 1], esto es, impulsar los logros humanos y organizativos mediante programas estratégicos de desarrollo de software para la sistematización exitosa de los procesos de sus clientes.

Microsoft ha experimentado un nuevo auge en 2023 gracias a la integración de la IA a sus productos [2]. Aquella empresa que fundara en abril de 1975, Bill Gates y Paul Allen en un pequeño local de Albuquerque, Nuevo México, es hoy la segunda, y ha sido también la primera, empresa más valiosa del mundo [3].

Cuenta con múltiples productos basados en IA, como lo es Azure OpenAI Service, una plataforma de aprendizaje automático que ofrece herramientas y servicios para que las empresas desarrollen e implementen sus propias soluciones en la nube. La plataforma incluye una amplia gama de soluciones como el procesamiento del lenguaje natural, la visión por computadora y la analítica predictiva; esta última se basa en el uso de datos históricos combinados con algoritmos estadísticos y técnicas de machine learning para identificar la probabilidad de resultados futuros [4]. También destaca por su asociación en 2023 con la empresa OpenAI [2], la imbricación de ChatGPT a su plataforma ofimática en Microsoft 365 Copilot y, Chat Bing como modelo de lenguaje generativo.

Microsoft y la IA responsable

En el ámbito de la ética de la IA, Microsoft ha publicado diversos documentos en los que expone sus principios y directrices para el desarrollo y uso de la inteligencia artificial. Estos documentos incluyen los Principios de IA y el sitio web IA responsable de Microsoft (Microsoft Responsible AI) [1] en los que la empresa expone que sus principios abarcan la equidad, la seguridad, la privacidad, la inclusión, la transparencia y la responsabilidad. Satya Nadella, CEO de Microsoft desde 2014, se ha comprometido a garantizar que la IA sea justa, accesible, segura y privada, y

que beneficie a la sociedad en su conjunto. También pretende crear una IA que sea inclusiva y fácil de usar, transparente sobre su funcionamiento y responsable de su impacto en la sociedad y el medio ambiente [1].

Para lo anterior, Microsoft cuenta con tres instancias en las que trabajan en la aplicación de los principios deontológicos de la empresa: la Oficina de IA responsable (ORA), el Comité Aether y la Estrategia de IA responsable en ingeniería (RAISE) [5].

La "Oficina de AI Responsable" tiene una labor de gobierno y política de Microsoft para una IA comprometida con sus principios dentro de la organización. Realiza cuatro tareas principales: a) Establece normas y directrices definiendo las responsabilidades y funciones en los equipos de IA en Microsoft; b) Capacita equipos para mejores prácticas de IA; c) Revisa los casos de uso sensibles; y d) Delimita las políticas públicas a través de leyes y normas [5].

El Comité Aether, por su parte, es un equipo de investigación y desarrollo que asesora a los directivos sobre riesgos y oportunidades emergentes en el campo de la IA; también realiza estudios, reflexiones y recomendaciones en respuesta a los temas que surgen en la empresa; y finalmente se encarga de las agendas de inclusión, equidad, confianza, transparencia, privacidad, seguridad e interacción entre la IA y las personas [5].

RAISE, por su parte, es un equipo de ingenieros dedicados a implementar las normas a los procesos de IA, crean los sistemas responsables y están desarrollando un sistema denominado Un sistema de ingeniería (One Engineering System - 1ES) para la estandarización de las normas en Microsoft, clientes y asociados; también trabajan en colaboración con ORA y Aether para identificar amenazas e implementar mejores prácticas [6].

Los principios específicos para una ética de la IA de Microsoft son:

1. Equidad

La importancia de la equidad en los sistemas de IA garantiza que no reivindiquen ni agraven los prejuicios y desigualdades sociales. La justicia penal, el empleo, la contratación, las finanzas y el crédito pueden reforzar los estereotipos existentes y llevar a una representación excesiva o insuficiente de algunos grupos sociales [1].

Según Microsoft, la equidad no es solo un reto técnico, sino un desafío sociotécnico que requiere un equipo que reflexione sobre los prejuicios y los aborden a lo largo de todo el ciclo de vida de la IA [7].

2. Confiabilidad y seguridad

La fiabilidad y la seguridad en todos los sistemas de IA debe garantizar que no causen daños a los usuarios. Si bien es posible que pueda haber errores, los riesgos deben cuantificarse y comunicarse a los afectados [7].

Este principio, según Microsoft, se aplica a todos los productos de IA, incluidos los modelos de aprendizaje automático (ML) que predicen si pueden causar daño directo a una persona; por ejemplo, los que se utilizan en servicios de salud o en conducción autónoma. La fiabilidad es una preocupación constante en todos los sistemas de IA, ya que pequeños errores pueden magnificarse si los sistemas defectuosos son utilizados a gran escala [8].

3. Privacidad

Microsoft se ha comprometido con la privacidad y la seguridad en sus sistemas y productos, y esto se extiende a la IA y el aprendizaje automático. Con el aumento de la dependencia de los datos para desarrollar y entrenar sistemas, deben existir nuevos requisitos para mantener el sistema a salvo de fallas y vulnerabilidades [1] [9].

La empresa aborda estas cuestiones ejecutando los modelos localmente en dispositivos de detección de anomalías para observar cambios en los datos que puedan indicar que un atacante está intentando penetrar en el sistema. La privacidad es un derecho fundamental y Microsoft asegura que toma medidas para garantizar que la información de usuarios gratuitos y clientes de pago estén protegidos frente a filtraciones o divulgaciones no autorizadas [1].

En diciembre de 2019 Microsoft confesó que millones de cuentas de Hotmail y Outlook habían sido vulneradas, pero argumentaron que esto se debió, principalmente, a un error humano y no del sistema en sí, porque los accesos de un empleado le fueron robados [10].

4. Inclusión y diversidad

Con el fin de empoderar e involucrar a las comunidades de todo el mundo, Microsoft pretende ser intencionadamente inclusivo y diverso con sus enfoques de la IA [1]. Quiere que todas las comunidades estén representadas en el diseño y la planificación de sus productos. Se subraya la importancia de garantizar que las experiencias sean realmente inclusivas; por ejemplo, que ningún idioma quede fuera. La ofimática Microsoft 365 y el navegador Microsoft Edge admiten en el año 2023, 86 idiomas [11].

5. Transparencia

La transparencia y la inteligibilidad son cruciales para alcanzar diversos objetivos, como mitigar la injusticia en los sistemas de aprendizaje automático, ayudar a los desarrolladores a depurar sus sistemas de IA y ganarse la confianza de los usuarios [1].

La transparencia tiene dos aspectos para Microsoft: en primer lugar, los creadores de sistemas deben ser abiertos sobre cómo y por qué utilizan la IA, y también sobre las limitaciones de sus algoritmos; en segundo lugar, los usuarios deben poder comprender el comportamiento de la IA, lo que se denomina interpretabilidad o inteligibilidad [14].

Por supuesto que el secreto comercial está presente. Windows no es un sistema operativo de código abierto, pero la empresa procura que sus algoritmos manejados por IA no se conviertan en una caja negra y que ni siquiera sus propios creadores sepan lo que sucede al interior. Para ello, cuenta con un Programa de seguridad gubernamental de Microsoft (Microsoft Government Security Program-GSP) con cinco Centros de transparencia alrededor del mundo: Estados Unidos, Bélgica, Singapur, Brasil y China. Estos centros están diseñados para aumentar la confianza en las ofertas de Microsoft y brindar a las agencias participantes la oportunidad de visitar una instalación de Microsoft y revisar su código fuente, así como otra información técnica relacionada con la seguridad [12] [15].

6. Responsabilidad

A pesar de la complejidad e imprevisibilidad de los nuevos modelos es importante rendir cuentas sobre el impacto de la tecnología en el mundo [1]. Microsoft considera que sus principios se promulguen de forma coherente y se tengan en cuenta en todas sus acciones y aplicaciones [1] [16].

Microsoft, en su documento "Estándar de IA responsable de Microsoft, v2" (Microsoft Responsible AI Standard, v2) [73] define la inteligencia artificial responsable como el desarrollo y la implementación de sistemas que ponen como prioridad consideraciones éticas como la transparencia, la equidad, la confiabilidad, la seguridad y la privacidad. La IA responsable tiene como objetivo garantizar que las tecnologías se desarrollen y utilicen de manera que beneficien a las personas, las organizaciones y la sociedad en su conjunto, al tiempo que se minimizan los riesgos potenciales e impactos negativos.

Discusión

De acuerdo con Microsoft, sus principios de IA dan prioridad a la equidad, la accesibilidad y minimizan los prejuicios en sus sistemas inteligentes con el objetivo de mejorar las capacidades humanas y beneficiar a la sociedad en su conjunto. Para ello, Microsoft se ha comprometido a crear una IA que sea inclusiva y fácil de usar, independientemente de la persona. Además, como se ha visto, hacen hincapié en la seguridad y la privacidad, garantizando que los datos se utilicen de forma responsable y transparente y que los sistemas estén protegidos frente a las ciberamenazas [17].

La transparencia y la responsabilidad son también principios fundamentales del desarrollo y el uso de la IA por parte de Microsoft. Su objetivo es proteger el código y entender cómo funcionan sus sistemas, qué datos utilizan y cualquier limitación o sesgo potencial de éste. Además, asumen la responsabilidad del impacto de sus productos en la sociedad y el medio ambiente trabajando para minimizar cualquier efecto negativo para crear una IA responsable. En general, los principios de la IA de Microsoft reflejan un enfoque integral y completo del desarrollo y el uso de su propia tecnología.

Se debe mencionar que Microsoft aprendió de su propio pasado y después de haber enfrentado entre 1990 y 1998 graves demandas monopólicas del Departamento de Justicia de los Estados Unidos por la Ley Sherman de defensa de la competencia [18], lo cierto es que los equipos ORA, Aether y RAISE están trabajando para crear productos y servicios que respeten no solo sus propios códigos de ética, sino las nuevas reglas del mercado global digital.

Google (Alphabet)

Aquella empresa fundada en septiembre de 1998 por Larry Page, Serguéi Brin y Scott Hassan [19] ha generado una dinámica de crecimiento exponencial, como

su propio nombre gúgol, que significa el número diez elevado a la centésima potencia [20]. Hoy Alphabet, dirigida por Sundar Pichai, CEO desde 2015, es un conjunto de empresas, y la más importante y conocida es Google, con sede en Mountain View, California.

La amplia base de datos de Google contiene información de más de 2 000 millones de usuarios de Internet y de teléfonos móviles de todo el mundo, como su ubicación geográfica, edad, sexo, búsquedas en Internet e intereses [22]. Google utiliza esta información para ofrecer búsquedas y publicidad personalizada a empresas de todo el mundo. Su mayor fuente de ingreso es Google Ads, que es la publicidad inserta en cada búsqueda [23]. Aunque este mercado está hoy en riesgo con los nuevos buscadores inteligentes como Perplexity AI, Andi Search o Wolfram Alpha, que no arrojan publicidad, sigue siendo Google el buscador más utilizado y la página de inicio de 9 de cada 10 computadoras en el mundo, según StatCounter.

Google usa su amplia base de datos para desarrollar otros productos y servicios. Por ejemplo, Google Maps, Google Translate y Google Play. Además, Google usa su base de datos para personalizar la experiencia de los usuarios, proporcionando resultados y anuncios de acuerdo con el perfil creado [24].

En cuanto a la ética, Google ha publicado diversos documentos que describen sus principios y pautas, incluido el documento “Inteligencia artificial en Google: nuestros principios” (AI at Google: our principles) [25] en el que contempla que la IA sea beneficiosa, sin prejuicios, responsable, privada, científica y no intrusiva. Google revisa cada año desde el 2019 estos principios a través de su *white paper* “Actualización sobre el progreso de los principios de IA” (AI Principles Progress Update) en los que describe los compromisos de Google con la transparencia, la seguridad, la equidad y la responsabilidad en el desarrollo y uso de la IA [26].

Otros temas más amplios que abarca la política de Google son: accesibilidad; arte y cultura; educación cívica; bienestar digital; diversidad e inclusión; educación de emprendedores; organizaciones sin fines de lucro; política pública; pequeñas empresas; programas para estudiantes y sostenibilidad e iniciativas como Google News; Google Org y Grow with Google [27]. Google también da prioridad a la investigación y la colaboración con la comunidad científica en general. Cada uno de estos temas representa una división de trabajo dentro de la empresa.

Por otra parte, Google ha manifestado que evita utilizar la IA con fines de vigilancia, a menos que exista un claro beneficio social [28]. Sin embargo, en 2018, por presiones de sus propios empleados, Google se vio en problemas por colaborar

con El Pentágono de los Estados Unidos en el Proyecto Maven de reconocimiento facial [29] [30]. Otro emprendimiento polémico fue el Proyecto Nimbus de colaboración con el gobierno de Israel, que en 2022 también despertó inquietudes por la aplicación del software Vision API, capaz de saber si alguien está mintiendo o no, a partir de la lectura de sus emociones en rostro por IA [31].

Sin embargo, para Google, en sus propias palabras, la IA representa una oportunidad para empoderar a las personas, beneficiar a las generaciones presentes y futuras y hacer avanzar el bien común [25]. Su propósito es organizar la información mundial y hacerla ampliamente disponible y útil. Google afirma estar consciente de que la tecnología fomenta la innovación y que se requiere adoptar ciertos principios para el desarrollo responsable que deben identificarse y aplicarse en las áreas concretas tanto de I+D como en la sociedad [25].

Por ejemplo, DeepMind, a cargo de Demis Hassabis es otra empresa filial de Google, que básicamente se dedica al desarrollo de nuevas aplicaciones de inteligencia artificial, como AlphaFold, que es un programa informático que predice la estructura tridimensional de las proteínas a partir de su secuencia de aminoácidos [21], entre otros proyectos de vanguardia.

Principios éticos propuestos por Google

1. La IA debe ser socialmente beneficiosa

El impacto de las nuevas tecnologías en la sociedad es cada vez mayor. Para Google diversos ámbitos, como la salud, la seguridad, la energía, el transporte, la manufactura y el ocio, pueden sufrir cambios radicales como consecuencia de los avances de la IA.

En Google consideran una amplia gama de variables sociales y económicas a la hora de evaluar el posible desarrollo y aplicaciones de su IA. Aseguran que, si las ventajas generales probables superan los peligros e inconvenientes previstos, siguen adelante, de lo contrario, afirman, se detienen [25].

2. Evitar crear o reforzar prejuicios

Los prejuicios injustos pueden verse reflejados, reforzados o atenuados por algoritmos y conjuntos de datos de IA. Google reconoce que no siempre es fácil distinguir los prejuicios, pero su objetivo es eliminar los efectos injustos sobre las personas, especialmente los que afectan a rasgos como la raza, etnia, sexo,

ingresos, nacionalidad, capacidades diferentes, orientación sexual, creencias políticas o religiosas [25].

3. IA construida con probada seguridad

Para evitar resultados imprevistos que aumenten riesgos, por ejemplo, de lesionar a alguien, Google afirma que, de acuerdo con los mejores métodos, aplica estrictos procedimientos de seguridad y protección para construir con precaución sistemas de IA. Evaluar las tecnologías en entornos limitados y vigilar su rendimiento una vez puestas en uso son parte de la sistematización de sus productos [25] [72].

4. Ser responsable ante las personas

Google considera que se deben construir sistemas que den la oportunidad a los usuarios para comentarios, explicaciones y apelaciones, por ello la IA desarrollada por la empresa debe estar supervisada y controlada por humanos, y en caso necesario, ser receptivos a la crítica para responder con criterios de personas y no de máquinas [25].

5. Diseño de privacidad

Los principios de privacidad deben observarse en la creación y aplicación de tecnología de IA, dice Google, por ejemplo, ofrecer la posibilidad de notificación y consentimiento. Debe ser una prioridad de los diseños que tengan en cuenta la transparencia y el control adecuado sobre el uso de los datos [25].

Por supuesto, Google ha enfrentado algunas crisis en este punto, pues con 2 000 millones de usuarios aproximadamente, es muy probable que existan violaciones a la privacidad, no siempre por culpa de la empresa, sino de los propios consumidores que usan contraseñas fácilmente reconocibles. Sin embargo, su sistema inteligente de Google Ads, capaz de generar campañas publicitarias personalizadas, ha sido cuestionado precisamente por el posible conflicto de la privacidad al rastrear digitalmente a los internautas. A su favor, hay que decir que Google permite la desactivación de anuncios personalizados, cookies y otros rastreadores desde su panel de control de herramientas de privacidad [24].

6. Excelencia científica

Para Google, el método científico, el trabajo colaborativo, la investigación abierta y el rigor intelectual son los cimientos de la innovación tecnológica. En campos

como la biología, la química, la medicina y las ciencias medioambientales, las herramientas de IA tienen el potencial de abrir nuevas perspectivas de investigación y comprensión epistémicas [25].

Google considera que debe existir colaboración con todas las partes interesadas, enfoques multidisciplinares y científicamente rigurosos, difusión de material de alta calidad, mejores prácticas del sector e investigaciones que permitan a más personas crear aplicaciones beneficiosas de IA [34].

7. Detener aplicaciones peligrosas o abusivas

Según Alphabet, para garantizar el desarrollo de una IA beneficiosa para todos, es importante tener en cuenta ciertos factores. En primer lugar, deben restringirse las tecnologías con potencial para un uso peligroso o abusivo. Esto implica evaluar si el propósito principal y la aplicación de una tecnología puede restringirse para evitar riesgos innecesarios [25] [35].

Por otra parte, Google advierte cuáles son los propósitos de la IA que no deben buscarse, por ejemplo: tecnologías que causen daño material, armas autónomas diseñadas para lesionar o matar y, tecnologías de vigilancia que violen el derecho internacional y los derechos humanos [25].

Google también creó un Código de conducta de proveedores de Google [36] para contribuir a un entorno de confianza y transparencia en su mercado. Está construido alrededor de los valores como respeto, diversidad e inclusión, y prohíbe cualquier tipo de acoso, abuso, castigo corporal o trato inhumano en cualquier ámbito de la cadena de producción en las empresas de sus asociados. El código también establece los requisitos para que los proveedores cumplan con los estándares ambientales [36].

Discusión

Google ha enfrentado con éxito una serie de escándalos desde 2013 con aquella campaña de Microsoft “Don’t get Scroogled” (No te dejes engañar) que advertía a los usuarios sobre la supuesta práctica de utilizar los datos personales con fines publicitarios sin su conocimiento o consentimiento [37].

Por otro lado, “Don’t be evil” (No seas malo) fue un eslogan informal y un código de conducta corporativo de Google creado por Paul Buccheit y Amit Patel [38], dos empleados destacados. La frase representaba el compromiso de promover una

cultura corporativa que diera prioridad a los beneficios a largo plazo y a la satisfacción de los usuarios frente a los beneficios a corto plazo. Rápidamente se convirtió en parte de la identidad de Google tanto interna como externamente. Sin embargo, en 2018, Google eliminó la cláusula "Don't be evil" de su código de conducta [39]. Aunque la frase ya no forma parte oficialmente de los valores de Google, sigue representando una visión importante de la historia y la ética de la empresa.

Google tiene un conjunto de principios que guían el desarrollo y uso de la IA en su investigación. Estos preceptos incluyen hacer que la información precisa y de alta calidad esté fácilmente disponible utilizando IA, sin dejar de respetar las normas culturales, sociales y legales en los países donde opera. Google también se ha comprometido a no utilizar su tecnología de IA para construir armas o realizar vigilancia ilegal.

Google afirma que utiliza un proceso de revisión para determinar si las nuevas tecnologías están en línea con sus principios y fomenta la capacitación de los responsables de la toma de decisiones y desarrolladores de IA para que tengan en cuenta las consideraciones éticas aquí esbozadas.

Facebook (Meta)

"Move Fast and Break Things" (Muévete rápido y rompe cosas) fue una bandera que tomó Mark Zuckerberg en 2010, cuyo objetivo era fomentar una cultura de innovación y experimentación en la empresa. Sin embargo, la frase también fue duramente criticada por promover un comportamiento, para muchos, imprudente y una falta de responsabilidad en las consecuencias que podía traer simplemente romper las cosas [40].

Facebook ha sido el villano de la película sobre ética corporativa. Sin embargo, Facebook (ahora Meta) ha invertido mucho en IA y aprendizaje automático para erradicar la incitación al odio y la desinformación en su plataforma, que han sido las principales acusaciones que se le han imputado [41]. Facebook también se ha enfrentado a preguntas sobre las implicaciones éticas de sus algoritmos y ha sido señalada de poner sus intereses comerciales por encima de sus propios principios éticos [42]. La empresa ha sido también criticada por el papel de sus algoritmos en la difusión de sesgos, información falsa, bullying, prejuicios raciales, políticos y de género [43] [45] [46] [47]

Sin embargo, Facebook utiliza la IA para identificar las publicaciones que podrían infringir sus normas antes de que las revisen los moderadores humanos [48], pero

al mismo tiempo, el propio algoritmo, según declaraciones de empleados, da prioridad a las publicaciones con mayor potencial de permanencia por parte de los usuarios, sin importar realmente lo que contengan sino las “reacciones”, visualizaciones y comentarios que estas provocan [44].

Sin embargo, hay que destacar que Meta ha publicado principios y directrices para el desarrollo y uso de la IA que abarcan seguridad y protección, rendición de cuentas, equidad, privacidad, investigación y colaboración. La empresa se ha comprometido a desarrollar una IA segura, robusta y fiable que sea transparente y explicable, así como responder a cualquier consecuencia negativa de sus servicios [41] [49].

Los pilares de Meta

Los pilares fundamentales de ética de la IA y de la tecnología que propone Meta son:

1. Privacidad y seguridad

Desarrollo de un proceso de revisión de la privacidad para evaluar los riesgos potenciales relacionados con la recopilación, el uso y el intercambio de la información de las personas [41].

La política de privacidad de Facebook incorpora diversas medidas de seguridad, como el cifrado de datos, para garantizar la protección de la información personal. En respuesta al escándalo de Cambridge Analytica, que se explicará en breve, Facebook ha dado un importante paso adelante. Han apostado por empoderar a los usuarios permitiéndoles controlar su configuración de privacidad, lo que les permite restringir la recopilación de sus datos por parte de terceros. Con una impresionante base de usuarios de casi 3 000 millones de personas, Facebook da prioridad a la salvaguarda de las identidades personales y a la seguridad general de la información de éstos [50].

2. Equidad e inclusión

Facebook defiende el principio de que sus productos deben ofrecer el mismo trato a todas las personas, sin discriminación. En consonancia con esta creencia, la imparcialidad se considera un aspecto clave de las expectativas de privacidad que rigen el proceso de revisión y las opciones disponibles para sus usuarios.

Al dar prioridad a la equidad, Facebook pretende garantizar que sus productos y servicios se diseñen e implementen de manera que traten a los usuarios de forma justa e imparcial. Esto significa que los algoritmos y mecanismos utilizados no deben favorecer ni discriminar a las personas en función de factores como la raza, el sexo, el origen étnico u otras características que deben ser protegidas [41].

3. Robustez y seguridad

Facebook reconoce la importancia de que los sistemas de IA funcionen de forma eficaz y fiable. Dan prioridad a la realización de pruebas exhaustivas para garantizar que estos sistemas puedan mantener su rendimiento de forma segura, incluso ante posibles ataques o amenazas. Para ello, ha creado un equipo especializado llamado Meta AI Research Topic (Tema de investigación de Meta IA). Este equipo se centra en evaluar la solidez de los sistemas de integridad frente a diversas amenazas potenciales. Al realizar estas evaluaciones, Facebook pretende mejorar la resistencia y la seguridad de sus sistemas, salvaguardando así la integridad de su plataforma y protegiendo a los usuarios de actividades dañinas o maliciosas [41].

4. Transparencia y control

La empresa ha implementado siete expectativas básicas de privacidad para que los usuarios tengan transparencia y control sobre cómo se recopila y utiliza la información [41]. Estas normativas son:

4.1 Control sobre la privacidad

El usuario debe saber quién ve qué. Puede seleccionar los destinatarios de cada publicación utilizando la herramienta de selección de público [50].

4.2 Explicabilidad de uso de datos

Facebook emplea un enfoque estratégico en el que integra herramientas y recursos informativos directamente en la experiencia del usuario. Además, se puede acceder a los controles de los anuncios en la esquina superior derecha de cada uno. Estas medidas se han puesto en marcha para que los usuarios comprendan mejor cómo se utilizan sus datos y para ofrecerles un mayor control sobre los anuncios que ven.

4.3 Diseño de privacidad

Resguardo de la protección de datos con seguridad. Para fomentar la transparencia, su objetivo es ofrecer una visión más clara del proceso de toma de decisiones de los sistemas de IA que afectan a los usuarios. Al hacer estos procesos más comprensibles, Meta pretende educar a la gente sobre el nivel de control que tienen sobre estas decisiones. De este modo, los usuarios pueden tomar decisiones con conocimiento de causa y comprender mejor cómo utilizan sus datos los sistemas de IA [50].

4.4 Mantener segura la información

Meta destaca su compromiso permanente con la seguridad de las cuentas y la integración de medidas de seguridad en todos sus productos. Para proteger proactivamente a los usuarios, Meta emplea sofisticados sistemas de seguridad que operan millones de veces para identificar y neutralizar las amenazas antes de que puedan afectar a los consumidores. Una medida de seguridad notable es la autenticación de dos factores, que mejora la protección de las cuentas. Además de introducir una contraseña, los usuarios deben introducir un código o token único que cambia dinámicamente y al que solo se puede acceder desde sus dispositivos móviles personales [50].

4.5 Derecho a eliminar la propia información

Los usuarios tienen el control de lo que publican, derecho a editar y, sobre todo, si lo desean, descartar algún contenido, e incluso eliminar su cuenta definitivamente [50].

4.6 Mejoras constantes

Según Facebook, las interfaces, así como los procesos, deben ser constantemente mejoradas para que sean intuitivas para todos los usuarios. Una interfaz intuitiva significa que requiere un esfuerzo mental mínimo para realizar una tarea. Cuanto menor sea la carga cognitiva, más podrán centrarse los usuarios en realizar su tarea [50].

4.7 Ser consecuentes

Esto lo interpreta Facebook como probar rigurosamente los productos para garantizar la seguridad de los datos y realizar evaluaciones exhaustivas de privacidad. También incluye recabar opiniones de legisladores, reguladores y expertos en privacidad de todo el mundo sobre prácticas y normativas en materia de datos [50].

5. Responsabilidad y gobernanza

Los procesos de toma de decisiones algorítmicas deben ser confiables, por lo que, en caso de una decisión crítica que tenga implicaciones éticas, estas deben ser tomadas por humanos [41].

6. Colaboración en una IA responsable

Facebook apela que aún no existen normas judiciales ni procedimientos establecidos universales para regular la IA. Sin embargo, se deben identificar y abordar los posibles efectos negativos relacionados con la IA como prioridad de la industria tecnológica, sin dejar de desarrollar nuevos productos. La comunidad de investigación, ingenieros, políticos y grupos de defensa deben trabajar unidos para hacer que la evaluación del impacto de la IA funcione a gran escala, sobre la base de normas transparentes y razonables. Antes de que la justicia, la privacidad, la solidez y la transparencia de la IA se consagren en la legislación, se debe trabajar para establecer normas básicas realistas, apunta la multinacional [41].

Discusión

Después de que la consultora política británica Cambridge Analytica usara los datos obtenidos indebidamente de aproximadamente 86 millones de usuarios de Facebook sin su consentimiento para manipular a los votantes indecisos durante las elecciones presidenciales de Estados Unidos de 2016 y en otras elecciones, el escrutinio de las políticas y prácticas de privacidad de datos de Meta sigue cuestionada [52]. Amnistía Internacional ha calificado a Cambridge Analytica como la "punta del iceberg" cuando se trata de empresas que hacen un uso indebido de datos personales con fines políticos o comerciales [53]. El documental The Great Hack (El gran hackeo o Nada es privado en Netflix) de los directores Karim Amer y Jehane Noujaim narra dicha historia [51].

La decisión de cambios de políticas de privacidad de Facebook se debió a que Cambridge Analytica trabajó con ellos y violó toda norma ética [50]. La firma usó los datos que obtuvo de Facebook para crear perfiles psicográficos detallados de los usuarios y dirigirlos con publicidad política personalizada. Además, esparció deliberadamente noticias falsas en la red [51].

El caso continúa abierto hasta 2023 y, en un acuerdo extrajudicial, Marx Zuckerberg se ha comprometido a pagar 725 millones de dólares para reparar el daño [54]. A partir de lo anterior, Facebook desarrolló, dentro de su organización, equipos

dedicados e interdisciplinarios para cambiar las prácticas informáticas y hacerlas de manera más responsable: Responsible AI, FAIR (Facebook Artificial Intelligence Research), AML (Applied Machine Learning) y una plataforma llamada FBLearner Flow [55]; hoy todos estos esfuerzos se concentran en Meta AI [56].

Facebook ha establecido las llamadas Normas comunitarias, que tienen como objetivo mantener a los usuarios seguros y promover interacciones positivas en la plataforma. Estos estándares prohíben el discurso de odio, la intimidación, el acoso y otras formas de contenido dañino. Facebook afirma que hace cumplir estos estándares a través de una combinación de moderadores humanos y herramientas de IA, así los usuarios pueden denunciar contenido que viole estos estándares [57].

Otro aspecto que se debe señalar, y que no se relaciona con IA, pero en los que Meta ha trabajado para reparar su imagen, ha sido sus acciones filantrópicas. Facebook tiene un historial de donaciones benéficas a través de la Fundación Chan Zuckerberg. La iniciativa apoya causas como la educación, la salud y la investigación científica [58]. Facebook también alienta a sus empleados a participar en el trabajo voluntario y les brinda oportunidades para hacerlo. La Iniciativa es una organización sin fines de lucro fundada por Zuckerberg y su esposa Priscilla Chan en 2015. La organización tiene como objetivo ayudar a resolver algunos de los desafíos más urgentes de la sociedad [58].

Hay que reconocer que Facebook se ha esforzado por aumentar la transparencia en torno a sus prácticas de datos y políticas publicitarias. Por ejemplo, la empresa ahora requiere que los anunciantes políticos verifiquen su identidad y ubicación, y proporciona una base de datos de búsqueda pública de todos los anuncios políticos en la plataforma. Facebook también ha hecho del dominio público un informe de transparencia que detalla las solicitudes gubernamentales de eliminación de datos y contenido de los usuarios [59].

Facebook últimamente se ha movido rápido, primero con su apuesta, aún en ciernes, por el Metaverso y ahora, tratando de recuperarse, a través de nuevas aplicaciones de inteligencia artificial [45]. La empresa ha enfrentado críticas y controversias sobre sus prácticas informáticas y comerciales, pero también cabe destacar los pasos positivos que ha tomado para abordar estos problemas y mejorar sus estándares.

Apple

Apple es en 2023 la empresa mejor cotizada del mundo y fue fundada por Steve Jobs, Steve Wozniak y Ronald Wayne en abril de 1976. Para la empresa de Cupertino, California, la privacidad de los datos de sus usuarios y la seguridad en general de sus sistemas operativos y dispositivos han sido una prioridad permanentemente. De hecho, están en el centro de la atención por su polémico sistema de bloqueo de publicidad personalizada, basada en el rastreo de navegación por aplicaciones de terceros [60]. En iOS 15, Apple introdujo un aviso que pregunta a los usuarios si quieren activar los anuncios personalizados, que antes estaban activados por defecto. Esto ha desatado una controversia porque se teme que la medida deje fuera del negocio a muchos de sus asociados y afiliados [61].

Apple ha publicado principios y directrices específicos para el desarrollo y uso de la tecnología [62] [63]. La empresa ha destacado la importancia de la privacidad y ha adoptado una postura firme en cuanto a la protección de los datos de los usuarios [64].

El CEO de Apple desde 2011 es Tim Cook, quien ha sido enfático sobre su compromiso ético, afirmando que la privacidad es un derecho humano fundamental [65]. En la práctica, esto se ha traducido en que Apple ha implementado funciones que impiden a terceros comercializar abiertamente las preferencias de los usuarios [60].

Se considera que la atención prestada por Apple a la privacidad es un paso positivo, sobre todo teniendo en cuenta la creciente preocupación y la seguridad de los datos. Algunos críticos, sin embargo, han argumentado que el ecosistema cerrado de Apple y las restricciones a las aplicaciones de terceros limitan la innovación y la competencia en sus sistemas [66].

Además, algunos han señalado que el énfasis de Apple en la privacidad a veces puede entrar en conflicto con otros objetivos, como mejorar la precisión de los algoritmos de IA, que a menudo se basan en grandes cantidades de datos [67]. Apple no tiene una política de ética sobre IA específica, pero sus principios corporativos aplican para todas sus áreas. El enfoque de Apple hacia la IA está más orientado a sus productos y tiende a centrarse en mostrar nuevas funciones habilitadas por IA, en lugar de discutir explícitamente los modelos y la tecnología de la IA. Dan, como ya se dijo, prioridad a la privacidad y la seguridad del usuario confiando en el procesamiento en el dispositivo y evitando la transferencia excesiva de datos a la nube.

En una larga entrevista [68], Cook habló de su preocupación por los posibles efectos negativos de la IA y de la importancia de dar prioridad a las consideraciones éticas en su desarrollo. Cook subraya que la IA tiene el potencial de ser una poderosa herramienta para el bien, pero que también presenta importantes retos para desincentivar el uso para el mal. Señaló que cuestiones como la parcialidad y el impacto de la automatización en los puestos de trabajo, la importancia de garantizar que la IA se desarrolle de forma transparente, responsable y respetuosa con los valores humanos, deben ser los pilares de desarrollo tecnológico.

Para hacer frente a estos retos, Cook y Apple han buscado la colaboración entre las empresas, los responsables políticos y el público. Piden un diálogo abierto y honesto sobre las implicaciones éticas de la IA y un compromiso compartido para desarrollarla de forma que beneficie a la sociedad.

De acuerdo con sus propios informes, Apple no recopila ni monetiza los datos de los usuarios del mismo modo que otras empresas tecnológicas, y las tecnologías de IA están diseñadas para funcionar en el dispositivo en lugar de depender del procesamiento basado en la nube [68].

Discusión

La empresa de la manzana también ha sido objeto de demandas, no precisamente por uso indebido de los datos, pero sí en otros temas éticos como la feroz obsolescencia programada [69] y prácticas presumiblemente monopólicas [66]. La obsolescencia programada es una estrategia empresarial en la que un producto se diseña y fabrica para que tenga intencionadamente una vida útil limitada o para que se quede obsoleto o menos funcional en un periodo de tiempo determinado. El objetivo es obligar a los consumidores a sustituir o actualizar sus productos con regularidad, creando así una demanda constante impuesta de nuevos productos y generando más ingresos cuando los dispositivos aún podrían servir [69].

Otros aspectos han tenido que ver con encerrar a sus socios y proveedores en un sistema operativo bloqueado y que, en algunos aspectos, raya en lo monopólico. Cuando los vendedores y afiliados están encerrados en un sistema operativo como macOS e iOS de Apple, significa que están restringidos en términos de qué software y hardware pueden utilizar y desarrollar para ese sistema [67].

MacOS es un sistema cerrado, lo que significa que está diseñado para funcionar únicamente con software y hardware aprobados por ellos mismos. Esto contrasta con un sistema abierto como Linux, que permite una mayor flexibilidad en términos

de qué software y hardware se puede utilizar, incluso Windows es menos restrictivo para los desarrolladores. Sin embargo, un sistema cerrado como macOS proporciona un alto nivel de seguridad, consistencia y coherencia a sus usuarios.

Conclusiones parciales

Existen al menos veinte laboratorios que están desarrollando inteligencia artificial de primer nivel en el mundo [70]; algunos están trabajando con las grandes empresas de manera coordinada, sin embargo, las directrices éticas siguen sin poderse delimitar y mucho menos alinear.

Microsoft ha publicado un conjunto de principios y directrices para el desarrollo y uso de la IA que se centran en la equidad, la seguridad, la privacidad, la inclusión, la transparencia y la responsabilidad.

Por su parte, Google (Alphabet) ha publicado principios y directrices para el desarrollo y uso de la IA que buscan que sea beneficiosa, sin prejuicios, responsable, privada, científica y no intrusiva.

Facebook (Meta), también tiene principios y directrices para el desarrollo y el uso de la IA que abarcan la seguridad, la responsabilidad, la equidad, la privacidad, la investigación y la colaboración. La literatura especializada considera que los principios y directrices de Meta, en materia de IA, son exhaustivos y están bien pensados, aunque algunos críticos sostienen que la empresa no ha estado en la práctica a la altura de los ideales que pregona.

Finalmente, Apple no ha publicado principios y directrices específicos para el desarrollo y uso de la IA, pero la empresa ha enfatizado la importancia de la privacidad y la protección de los datos de los usuarios. Se considera positivo que Apple se centre en la privacidad, aunque algunos críticos han señalado que a veces puede entrar en conflicto con otros objetivos, como mejorar la precisión de los algoritmos de IA y dar espacio a otros para la libre competencia dentro de sus sistemas.

Se concluye que los códigos de ética pueden ser una forma eficaz de promover prácticas responsables y crear estándares morales, tanto en las propias empresas como en sus afiliados, mientras que otros se muestran más escépticos y los consideran un tipo de “lavado de cara” ético o ejercicios de relaciones públicas, carentes de un correlato objetivo en sus prácticas corporativas. Por ejemplo, refieren a las condiciones de trabajo en manufactureras chinas que trabajan para Apple [74].

Una de las principales críticas a las normas de auto regulación es que son voluntarias y carecen de mecanismos de aplicación, lo que dificulta exigir responsabilidades a las empresas si no cumplen sus propios principios declarados; algunos, debe señalarse, cuentan con equipos especializados al interior que se dedican presumiblemente a garantizar que esos códigos se cumplan.

Los escépticos también señalan que, algunas empresas pueden utilizar las normas de auto regulación como una forma de desviar las críticas y evitar un reglamento gubernamental más estricto, toda vez que al público se muestran explícitamente partidarios de seguir ciertos lineamientos [71].

Por otra parte, los defensores de las normas autoimpuestas sostienen que pueden ser una herramienta importante para promover la transparencia, la rendición de cuentas y las prácticas responsables en el desarrollo y el uso de la IA. Las empresas pueden utilizar, por ejemplo, “códigos de honor” para señalar su compromiso con los principios éticos y entablar un diálogo con las partes interesadas, incluidos clientes, empleados y grupos de la sociedad civil defensores de una IA responsable, pero no se garantiza en ningún momento su cumplimiento.

Por tanto, y en última instancia, la eficacia de las normas auto reguladoras dependerá de una serie de factores, como la voluntad de la empresa por entablar un diálogo significativo con todas las partes, la solidez de estos principios y la capacidad de los usuarios —y los entes jurisdiccionales— para exigir que rindan cuentas de sus acciones.

En conclusión, las recientes iniciativas de ética de la IA de empresas tecnológicas mundiales como Microsoft, Google, Meta y Apple han atraído la atención sobre los retos éticos que plantea la IA y la importancia de un desarrollo e implantación responsables. Aunque estas iniciativas representan un paso adelante positivo, sigue habiendo dudas sobre su eficacia y su potencial de compromiso. Estas empresas tienen la responsabilidad de ir más allá de los cambios superficiales o las tácticas de relaciones públicas y dar prioridad realmente a los principios éticos y las prácticas responsables en una competencia abierta por el liderazgo de la IA a nivel global. Lo que no se puede negar es que aun cuando a las empresas las mueven los fines de lucro, todas ellas han contribuido, de una u otra manera, a que el mundo sea más hospitalario, menos ignorante e interconectado.

Referencias

[1] Microsoft, “Putting principles into practice at Microsoft,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ciU

[2] N.C.M. Latinoamérica, “Microsoft y OpenAI amplían su asociación,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://news.microsoft.com/es-xl/microsoft-y-openai-amplian-su-asociacion/

[3] Microsoft, “The History of Microsoft,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://learn.microsoft.com/en-us/shows/history/

[4] Microsoft, “¿Qué es Azure?,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3h

[5] Microsoft, “Ponemos en práctica nuestros principios,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ciV

[6] Microsoft, “Driving engineering culture change at Microsoft: An experimental journey,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3i

[7] Microsoft, “Responsible and trusted AI,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3j

[8] Microsoft, “Empowering impactful responsible AI practices,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ciW

[9] W. Pearce y R. Siva Kumar, “Best practices for AI security risk management,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3l

[10] M. Heiligenstein, “Microsoft Data Breaches: Full Timeline Through 2023,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://firewalltimes.com/microsoft-data-breach-timeline/.

[11] Microsoft, “International availability,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://www.microsoft.com/en-us/microsoft-365/business/international-availability.

[12] Microsoft, “Empowerment begins with trust [video],” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://www.microsoft.com/en-us/trust-center#modal1.

[13] Microsoft, “Mantener el control de la privacidad,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3p.

[14] Microsoft, “Transparency. AI systems should be understandable [video],” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ciX

[15] Microsoft, “Transparency Centers,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://learn.microsoft.com/en-us/security/gsp/contenttransparencycenters

[16] N. Crampton, “Microsoft’s framework for building AI systems responsibly,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3k.17

[17] T. C. King, N. Aggarwal, M. Taddeo, y L. Floridi, “Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions,” en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed. Cham: Springer International Publishing, 2021, 251-282.

[18] C. Team, “Microsoft Antitrust Case,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ciY

[19] Feedough, “The History of Google,” Feedough.com. Acceso jun. 2023. [En línea] Disponible: https://www.feedough.com/the-history-of-google/

[20] M. Bellis, “The History of Google and How It Was Invented,” 2021. Thoughtco.com. Acceso jun. 2023. [En línea] Disponible: https://www.thoughtco.com/who-invented-google-1991852

[21] S. Pappas, “What is DeepMind?,” Livescience.com. Acceso jun. 2023. [En línea] Disponible: https://www.livescience.com/what-is-deepmind

[22] Google, “Bringing an inclusive, equitable Internet to everyone, everywhere,” Google.com. Acceso jun. 2023. [En línea] Disponible: https://nextbillionusers.google/a-z/

[23] Wordstream, “Google Ads: What Are Google Ads & How Do They Work?,” Google.com. Acceso jun. 2023. [En línea] Disponible: https://www.wordstream.com/google-ads.

[24] Google, “Google Ads,” Google.com. Acceso jun. 2023. [En línea] Disponible: https://ads.google.com/

[25] Google, “Our Principles. Google AI,” Google.com. Acceso jun. 2023. [En línea] Disponible: https://ai.google/principles/

[26] Google, “AI Principles reviews and operations,” Google.com. Acceso jun. 2023. [En línea] Disponible: https://ai.google/responsibilities/review-process/

[27] Google, “About Google,” Google.com. Acceso jun. 2023. [En línea] Disponible: Disponible en: https://about.google/

[28] Google, “Nos comprometemos a mejorar significativamente la vida de la mayor cantidad de personas posible,” Google.com. Acceso jun. 2023. [En línea] Disponible: https://about.google/commitments/

[29] N.Y. Times, “The Business of War: Google Employees Protest Work for the Pentagon,” NYTimes.com. Acceso jun. 2023. [En línea] Disponible: https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html.

[30] K. Conger and D. Cameron, “Google Is Helping the Pentagon Build AI for Drones,” 2018. Gizmodo.com. Acceso jun. 2023. [En línea] Disponible: https://gizmodo.com/google-is-helping-the-pentagon-build-ai-for-drones-1823464533

[31] Silberling, “Google workers protest $1.2B Project Nimbus contract with Israeli military,” Techcrunch.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ciZ

[32] D. Acemoglu, D. Autor, J. Hazell y P. Restrepo, “Artificial Intelligence and Jobs: Evidence from Online Vacancies,” *J. Labor Econ*., vol. 40, 2022. https://bsu.buap.mx/ciR

[33] N. Tiku, “The Google engineer who thinks the company’s AI has come to life,” Washingtonpost.com. Acceso jun. 2023. [En línea] Disponible: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

[34] Google, “Google Research, 2022 & beyond: Research community engagement,” Google.com. Acceso jun. 2023. [En línea] Disponible: https://ai.googleblog.com/2023/02/google-research-2022-beyond-research.html.

[35] M. Taddeo, T. McCutcheon, and L. Floridi, "Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword," in *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed. Cham: Springer International Publishing, 2021, pp. 289-297, doi: 10.1007/978-3-030-81907-1_15.

[36] Google, "Código de Conducta de Proveedores de Google," Google.com. Acceso jun. 2023. [En línea] Disponible: https://about.google/supplier-code-of-conduct/

[37] Microsoft, "Don't Get Scroogled by Gmail," Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://news.microsoft.com/2013/02/07/dont-get-scroogled-by-gmail/

[38] CNET, "Don't be evil': Google's iconic mantra comes into question at labor trial," Cnet.com. Acceso jun. 2023. [En línea] Disponible: https://www.cnet.com/tech/tech-industry/dont-be-evil-googles-iconic-mantra-comes-into-question-at-labor-trial/.

[39] R. Lawler, "Alphabet replaces Google's 'Don't be evil' with 'Do the right thing'," Engadget.com. Acceso jun. 2023. [En línea] Disponible: https://www.engadget.com/2015-10-02-alphabet-do-the-right-thing.html

[40] J. Taplin, Move Fast and Break Things. Boston, MA, USA: Little, Brown and Company, 2017.

[41] MetaAI, "Facebook's five pillars of Responsible AI," Facebook.com. Acceso jun. 2023. [En línea] Disponible: https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/.

[42] D. Lauer, "Facebook's ethical failures are not accidental; they are part of the business model," AI and Ethics, vol. 1, no. 4, pp. 395-403, 2021, doi: 10.1007/s43681-021-00068-x

[43] T. Wayt, "History will not judge us kindly: Facebook employees rip Zuckerberg in leaked messages," NYPost.com. Acceso jun. 2023. [En línea] Disponible: https://nypost.com/2021/10/25/facebook-employees-flag-ethical-concerns-rip-zuckerberg/

[44] K. Paul y D. Milmo, "Facebook putting profit before public good, says whistleblower Frances Haugen," TheGuardian.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3q.

[45] MetaAI, "ImageBind: a new way to 'link' AI across the senses," Metademolab.com. Acceso jun. 2023. [En línea] Disponible: https://imagebind.metademolab.com/

[46] K. Hao, "How Facebook got addicted to spreading misinformation," Technologyreview.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ci0

[47] E. Dwoskin, "Facebook's Sandberg deflected blame for Capitol riot, but new evidence shows how platform played role," Washingtonpost.com. Acceso jun. 2023. [En línea] Disponible: https://www.washingtonpost.com/technology/2021/01/13/facebook-role-in-capitol-protest.

[48] M. Schroepfer, "AI gets better every day. Here's what that means for stopping hate speech," Facebook.com. Acceso jun. 2023. [En línea] Disponible: https://ai.facebook.com/blog/ai-gets-better-every-day-heres-what-that-means-for-stopping-hate-speech/

[49] M. Taddeo y L. Floridi, "The Debate on the Moral Responsibilities of Online Service Providers," *Sci. and Eng. Ethics*, vol. 22, no. 6, pp. 1575-1603, 2016, doi: 10.1007/s11948-015-9734-1.

[50] Meta, “Centro de Privacidad,” 2023. [En línea]. Disponible en: https://www.facebook.com/privacy/policy/.

[51] K. Amer y J. Noujaim, “The Great Hack [Nada es privado],” Netflix, Estados Unidos, 2019.

[52] BBC, “5 claves para entender el escándalo de Cambridge Analytica que hizo que Facebook perdiera US$37.000 millones en un día,” BBC.com. Acceso jun. 2023. [En línea] Disponible: https://www.bbc.com/mundo/noticias-43472797

[53] Internacional, “El gran hackeo: Cambridge Analytica es solo la punta del iceberg,” Amnesty.org. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ci1

[54] France24, “La matriz de Facebook acepta pagar USD 725 millones por escándalo de Cambridge Analytica,” France24.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3s

[55] K. Hao, “Cómo Facebook se volvió adicto a la desinformación,” Google.com. Acceso jun. 2023. [En línea] Disponible: https://radiolaprimerisima.com/como-facebook-se-volvio-adicto-a-la-desinformacion/

[56] Meta, “Inside the Lab: Building for the Metaverse With AI,” Meta.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ci2

[57] Meta, “Normas comunitarias de Facebook,” Meta.com. Acceso jun. 2023. [En línea] Disponible: https://transparency.fb.com/es-la/policies/community-standards/

[58] C.Z. Initiative, “The Chan Zuckerberg Initiative,” Chanzuckerberg.com. Acceso jun. 2023. [En línea] Disponible: https://chanzuckerberg.com/

[59] Meta, “Transparency reports,” Facebook.com. Acceso jun. 2023. [En línea] Disponible: https://transparency.fb.com/data/

[60] Apple, “Política de privacidad de Apple,” Apple.com. Acceso jun. 2023. [En línea] Disponible: https://www.apple.com/mx/legal/privacy/es-la/

[61] C. Mackintosh, “Privacy policies for iOS apps,” Harperjames.com. Acceso jun. 2023. [En línea] Disponible: https://harperjames.co.uk/article/apple-privacy-policy-apps/

[62] Apple, “Shared Values,” Apple.com. Acceso jun. 2023. [En línea] Disponible: https://www.apple.com/careers/us/shared-values.html.

[63] Apple, “Ethics and Compliance,” Apple.com. Acceso jun. 2023. [En línea] Disponible: https://www.apple.com/compliance/.

[64] A.J. Andreotta, N. Kirkham, y M. Rizzi, “AI, big data, and the future of consent,” *AI & Soc.*, vol. 37, no. 4, pp. 1715-1728, 2022, doi: 10.1007/s00146-021-01262-5

[65] NPR, “Apple CEO Tim Cook: ‘Privacy Is A Fundamental Human Right’,” NPR.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3w

[66] F. Yun Chee y S. Kar-Gupta, “EU antitrust regulators narrow charges against Apple,” Reuters.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ci3

[67] Tilley, “U.S. Escalates Apple Probe, Looks to Involve Antitrust Chief,” 2023. Google.com. Acceso jun. 2023. [En línea] Disponible: https://www.wsj.com/articles/u-s-escalates-apple-probe-looks-to-involve-antitrust-chief-2fa86ddf

[68] Z. Baron, “Tim Cook on Shaping the Future of Apple,” GQ.com. Acceso jun. 2023. [En línea] Disponible: https://www.gq.com/story/tim-cook-global-creativity-awards-cover-2023

[69] S. Keach, “Planned obsolescence: Apple set to ‘kill off’ older iPhone models,” NYPost.com. Acceso jun. 2023. [En línea] Disponible: https://nypost.com/2021/12/02/apple-may-soon-kill-off-older-iphone-models/

[70] Komarraju, “Top 20 Artificial Intelligence Research Labs In The World In 2021,” Analyticsinsight.com. Acceso jun. 2023. [En línea] Disponible: Disponible en: https://www.analyticsinsight.net/top-20-artificial-intelligence-research-labs-in-the-world-in-2021/

[71] T. Hagendorff, “The Ethics of AI Ethics: An Evaluation of Guidelines,” Mind and Mach., vol. 30, no. 1, pp. 99-120, 2020, doi: 10.1007/s11023-020-09517-8

[72] WPKube, “The Google Cemetery,” WPKube.com. Acceso jun. 2023. [En línea] Disponible: https://gcemetery.co/

[73] Microsoft, “Microsoft Responsible AI Standard, v2,” Microsoft.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ckS

[74] FoxNews, “Apple supply workers describe noxious hazards, unsafe conditions at China factory”, 2018, Acceso jul. 2023. [En línea] Disponible: https://bsu.buap.mx/cr4

ESTADOS UNIDOS, CHINA Y RUSIA: PROPUESTAS NACIONALES PARA UNA ÉTICA DE LA IA EN LA NUEVA GUERRA FRÍA

Introducción

Las normas nacionales y el campo de batalla tecnológico, industrial, económico e incluso militar de las aplicaciones de la inteligencia artificial se enfrentarán significativamente entre las potencias mundiales a medida que avance la carrera por la supremacía geopolítica estratégica. China y Estados Unidos están considerados como las dos naciones punteras en I+D (investigación y desarrollo). Estados Unidos alberga algunas de las principales empresas e instituciones de IA, y cuenta con un sólido historial de innovación y espíritu emprendedor en el sector tecnológico. China, por su parte, ha avanzado considerablemente en I+D gracias a la financiación pública, a un gran número de jóvenes investigadores y al espíritu de competencia del Partido Comunista Chino. Rusia, por su parte, con un liderazgo pragmático y demencial, avanza en IA a paso firme y sin mayores reservas. Como advirtió el presidente ruso, Vladímir Putin: “la nación que domine los avances en la tecnología de IA gobernará el mundo” [16]. Cada uno de los tres países ha publicado su normativa ética para la IA y, como objetivo de este capítulo, se presenta un análisis ponderado de cada uno.

Los Estados Unidos de Norteamérica

Estados Unidos lanzó la “Ley de iniciativa nacional de inteligencia artificial de 2020” (NAIIA - National AI Initiative Act of 2020, Division E, Sec. 5001) [1], que se convirtió en ley a partir del 1° de enero de 2021 [1]. Es un proyecto aprobado por el Congreso que esboza un conjunto de políticas destinadas a impulsar la investigación y el desarrollo en el campo de la IA. Uno de los principales objetivos de esta ley es aumentar el liderazgo y la capacidad de investigación de Estados Unidos en la materia.

Para alcanzar este objetivo, la ley propone una serie de medidas para promover la colaboración y la coordinación entre los diferentes sectores y partes interesadas, incluidas las agencias gubernamentales, las instituciones académicas, la industria privada y los socios internacionales. La ley crea la "Oficina de la iniciativa nacional de inteligencia artificia" (NAIIO - The National Artificial Intelligence Initiative Office) [2] para supervisar y coordinar estos esfuerzos con el mandato de promover la investigación, el desarrollo y el despliegue de tecnologías de IA que sirvan al interés público, pero sobre todo a los intereses del gobierno de los Estados Unidos [3].

De este modo, la ley autoriza a la NAIIO a preparar un plan estratégico de investigación y desarrollo de la IA. El plan se centra en el avance de las tecnologías básicas, el perfeccionamiento y la formación en conjunción con el trabajo, la garantía de un uso ético y responsable, y el apoyo a la innovación y la comercialización. La ley también financia una serie de programas e iniciativas de apoyo a la educación y el desarrollo de infraestructura de IA en diversos ámbitos [1].

Los propósitos de la ley son:

Asegurar el liderazgo de los Estados Unidos en investigación de la IA.

Liderar el mundo en el desarrollo y uso de sistemas de IA fiables en los sectores público y privado.

Preparar a la mano de obra actual y futura de ese país para la integración de los sistemas de IA en todos los sectores de la economía y la sociedad.

Coordinar las actividades de investigación, desarrollo y demostración de IA entre las agencias civiles, el Departamento de Defensa y la comunidad de inteligencia para asegurar que cada uno informa su trabajo a los demás [4].

Los seis pilares que contempla la ley gubernamental para la fortaleza de su política en IA son:

1. I+D investigación y desarrollo

Dar prioridad a la investigación y el desarrollo de IA. Uno de los principales objetivos de la ley es mantener el liderazgo y la capacidad de investigación de Estados Unidos en este campo. Para alcanzar esto, la ley propone una serie de medidas para la colaboración y la coordinación entre los diferentes sectores y partes interesadas,

incluidas las agencias gubernamentales, las instituciones académicas, la industria privada y los socios internacionales [1].

2. Infraestructura de investigación de IA

Otro de los objetivos de la ley es mejorar la calidad y accesibilidad de los datos, modelos y recursos informáticos; esto es, el acceso de los investigadores y desarrolladores de IA a insumos de alta calidad. Esto lo esperan lograr mediante el establecimiento de asociaciones entre el gobierno, la industria privada y las instituciones académicas para facilitar la colaboración y el intercambio de datos [1].

También contempla que, para promover la investigación y el desarrollo de la IA, apoyar un crecimiento más equitativo, ampliar el conocimiento y permitir su aplicación a una mayor variedad de campos, se necesita aumentar el acceso a recursos computacionales de vanguardia y a conjuntos de datos de alta calidad, por lo que debe ser posible tener acceso a la infraestructura de la IA y, al mismo tiempo, conservar las garantías de confidencialidad, seguridad y privacidad [1].

3. IA avanzada y confiable

La ley contempla modernizar las normas de las tecnologías basadas en IA con el fin de proteger la privacidad, los derechos civiles, las libertades públicas y otros valores democráticos; un esfuerzo coordinado para establecer normas técnicas y marcos de gobernanza que garanticen el desarrollo y despliegue responsables de la IA [1].

Lo anterior, también significa proteger la privacidad y las libertades civiles en el desarrollo y uso de las tecnologías mediante el establecimiento de marcos de privacidad y seguridad de los datos, así como el desarrollo de mecanismos de transparencia y rendición de cuentas. La ley, con el objetivo de promover prácticas éticas y responsables, aboga por la permanente actualización del marco jurídico y normativo que regule el uso de la IA [5].

4. IA para la gobernanza y la seguridad nacional

Este pilar se refiere a aplicar tecnologías de IA para la seguridad nacional y la prestación de servicios públicos de calidad en Estados Unidos. Observar la eficiencia, la eficacia y la capacidad de respuesta ante amenazas internas y externas [1].

La NAIIA pretende abogar el desarrollo de nuevas tecnologías basadas en la IA para la prestación de servicios públicos como salud, transporte y educación [6]. Mejorar la seguridad nacional implica el uso de la IA en materia de ciberseguridad, inteligencia estratégica y capacidades físicas de defensa. Por ejemplo, tecnologías habilitadas para IA, como los drones autónomos o los sistemas de vigilancia, pueden mejorar la conciencia situacional, el control territorial y la protección de infraestructuras críticas [7].

Por otra parte, la prestación de servicios públicos mediante el uso de la IA implica el desarrollo y despliegue de tecnologías que mejoren la eficiencia, eficacia y capacidad de respuesta a la ciudadanía. Por ejemplo, la ley prevé el uso de la IA en la salud para optimizar el diagnóstico, el tratamiento y la prevención de enfermedades, así como en el transporte, mejorar la fluidez del tráfico y reducir los accidentes [8].

5. Compromiso internacional

Promover el compromiso internacional para el desarrollo de IA tiene como propósito crear un entorno global que apoye los valores democráticos de Occidente. El objetivo principal es establecer asociaciones con aliados de ideologías afines, especialmente el Reino Unido, para promover el desarrollo y el despliegue de tecnologías de forma que se ajusten a sus ideales y fomenten un entendimiento compartido de los beneficios y riesgos inherentes a la IA [1].

La NAIIA pretende alcanzar esto mediante el establecimiento de nuevas asociaciones con países aliados, así como con organizaciones y foros internacionales. Estas asociaciones se basan, de acuerdo con el gobierno, en el intercambio de ideas, mejores prácticas y conocimientos especializados relacionados con el desarrollo y la implantación de nuevas tecnologías [9].

Esto podría implicar el desarrollo de marcos y normas comunes para el despliegue responsable de sistemas de IA, así como la promoción de la transparencia y la rendición de cuentas en su uso. La ley también pide la creación de nuevas iniciativas para apoyar el desarrollo de tecnologías que promuevan el bien social, como el uso de la IA para abordar retos globales como el cambio climático, la pobreza y la desigualdad [10].

6. Fuerza laboral capacitada lista para la IA

Otro objetivo de la ley es proporcionar educación para prepararse para el uso de la IA en todos los niveles, desde el kínder hasta la universidad, pasando por la mano de obra versada en nuevas tecnologías [1].

Esta iniciativa se centra en el desarrollo de nuevos planes de estudios y programas de formación para proporcionar a los estudiantes los conocimientos y habilidades necesarios para tener éxito en una economía impulsada por la IA. Asimismo, la ley exige la creación de nuevas iniciativas de financiación para apoyar la investigación y el desarrollo en IA, así como nuevas asociaciones entre la industria y las instituciones académicas [11].

Con educación, Estados Unidos espera poner en marcha políticas que garanticen una mano de obra diversa, integradora y preparada, y cerrar así la brecha de cualificación en IA y preparar a los empleados estadounidenses para las ocupaciones del futuro [12] [13].

Discusión

La NAIIA busca establecer una estrategia nacional coordinada para el desarrollo y despliegue de las tecnologías de IA centrada en promover el crecimiento económico, reforzar la seguridad nacional y mejorar la calidad de vida de los estadounidenses. La ley establece una serie de requisitos y directrices que deben seguir los organismos federales para alcanzar estos objetivos, como el establecimiento de una oficina especializada (NAIIO), la creación de un Comité Asesor y una mayor colaboración entre el gobierno, el mundo académico y la industria.

Sin embargo, como cualquier legislación, la ley está sujeta a la interpretación y aplicación del Gobierno Federal. Es posible que haya áreas en las que el propio gobierno no esté cumpliendo plenamente la ley, o en las que existan diferencias de opinión sobre la mejor manera de alcanzar sus objetivos. También cabe señalar que la ley no aborda específicamente el uso de las tecnologías de IA en el campo de batalla, por lo que no está claro cómo se aplicaría para uso bélico en el contexto internacional o si no aplique [7].

La literatura especializada sugiere que la NAIIA es un paso importante para promover el desarrollo responsable y el despliegue de las tecnologías de IA en Estados Unidos. Sin embargo, dependerá de las agencias gubernamentales, la industria y otras partes interesadas trabajar juntos para garantizar que los objetivos

de la ley se alcancen de una manera que sea coherente con otras normativas éticas y legales aceptadas.

La Federación de Rusia

A partir de la invasión de Rusia a Ucrania en febrero de 2022, esta nación ha sido objeto de sanciones económicas y dura condena internacional por la guerra desplegada, especialmente por los ataques a la población civil considerados como crímenes de guerra y denunciados por los medios de comunicación y organismos internacionales [14]. Hoy continúan los combates que, probablemente, se definan por los drones kamikazes y otras armas inteligentes teledirigidas utilizadas por ambos bandos [15].

Vladimir Putin había dicho durante el Foro de navegación y proyección profesional de Yaroslavl en 2017, que: "la nación que domine los avances en la tecnología de inteligencia artificial gobernará el mundo" [16]. Eso es algo que Rusia se ha tomado en serio [17].

En octubre de 2021, Rusia organizó un foro titulado "Ética de la inteligencia artificial: el comienzo de la confianza" [18] en el que autoridades federales y regionales, empresarios y público representante de la sociedad civil discutieron el impacto de la IA en las áreas de la vida y los riesgos asociados con su uso.

De este evento surgió un documento denominado: "Código de ética para la inteligencia artificial" (Кодекс этики в сфере ИИ) [19], en el que se fijaron las siguientes disposiciones referentes a los diversos ámbitos de producción y aplicación ética de la IA. Los preceptos son los siguientes:

1. Intereses y derechos del individuo

Para la Federación Rusa, la prioridad del desarrollo de tecnologías de IA para proteger los intereses y derechos de las personas y del individuo son:

1.1 Un enfoque humanista

Los derechos y libertades del individuo deben tener el máximo valor a la hora de desarrollar tecnologías de IA. El potencial humano para alcanzar la armonía en los ámbitos social, económico y espiritual, así como la máxima realización personal, debe ser promovido y no obstaculizado por la tecnología alcanzada por la IA [20].

Rusia considera que la autonomía y la capacidad de decisión de un ser humano, la libertad de elección entre diversas alternativas y, en general, los talentos intelectuales deben ser preservados por los actores de la IA como un valor intrínseco y un elemento formador en la sociedad contemporánea. A la hora de desarrollar sistemas de IA, los actores deben tener en cuenta cualquier daño potencial para el crecimiento de las habilidades cognitivas humanas y tomar medidas para detener aquella IA que intencionadamente tenga efectos adversos [19].

Deben tenerse en cuenta principios importantes, como la preservación y el desarrollo de la creatividad y las capacidades cognitivas humanas, la defensa de los valores morales, espirituales y culturales, el fomento de la diversidad y la identidad cultural y lingüística, y la preservación de las costumbres y las raíces de las naciones, los pueblos y los grupos étnicos y sociales [20]. Rusia está conformada por 46 *oblasts* (provincias o estados) y 22 repúblicas, entre otras divisiones políticas.

1.3 Cumplimiento de la ley

Para entender y adherirse a la legislación de la Federación Rusa relativa a las tecnologías de IA, es importante familiarizarse con el denominado Concepto Regulador. Este concepto está diseñado para establecer las condiciones necesarias para normar las interacciones sociales emergentes que están asociadas con el desarrollo y la implementación de la IA. Su objetivo es proporcionar un marco para abordar las consideraciones jurídicas y éticas en torno a la tecnología de IA en Rusia. Aprobado por "Decreto del Gobierno de la Federación Rusa del 19 de agosto de 2020 No. 2129-r" (Концепция развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 года) [21], su objetivo es determinar enfoques para transformar el sistema regulatorio para garantizar la posibilidad de crear y aplicar IA y tecnologías robóticas en diversos sectores de la economía respetando los derechos de los ciudadanos y garantizando la seguridad del individuo, la sociedad y el Estado.

1.4. No discriminación

Los actores de la IA deben tomar precauciones para garantizar la equidad y la no discriminación, asegurándose de que los algoritmos de aprendizaje automático, los conjuntos de datos y las técnicas de procesamiento utilizadas para agrupar y categorizar información sobre individuos o grupos no discriminen deliberadamente a nadie. Se fomenta entre los actores de la IA el desarrollo y la aplicación de técnicas y herramientas de software que reconozcan y prohíban la discriminación

por motivos de raza, nacionalidad, género, creencias políticas, creencias religiosas, edad, posición social, posición económica o detalles de la vida privada de una persona [19].

1.5 Evaluación de riesgos e impacto humanitario

En algunas fases, incluso durante la creación y el uso de conjuntos de datos, se anima a los actores de la IA a evaluar los peligros potenciales, así como el impacto humanitario del sistema sobre los derechos humanos y las libertades. Las manifestaciones de estos peligros deben supervisarse a largo plazo y, al evaluar los riesgos, los agentes deben tener en cuenta la complejidad del comportamiento de los sistemas, incluida la interconexión y conectividad entre procesos a lo largo de los ciclos de vida [17].

Lo anterior, siempre y cuando no ponga en peligro la funcionalidad y la seguridad de la información del sistema y garantice la protección de la propiedad intelectual y los secretos comerciales del desarrollador, por lo que se recomienda realizar una evaluación de riesgos de las aplicaciones críticas con la ayuda de un agente que sea neutral o de un organismo oficial autorizado.

2. Ser consciente de la responsabilidad de crear y utilizar IA

2.1. Enfoque basado en el riesgo

La atención prestada a las preocupaciones éticas relativas a la IA y a las acciones emprendidas debe estar en consonancia con la evaluación de los riesgos asociados a las tecnologías y sistemas específicos de la IA en beneficio de las personas y la sociedad. Deben tenerse en cuenta tanto los peligros conocidos como los potenciales, incluida la probabilidad y el impacto previsible de las amenazas a corto y largo plazo. Las decisiones relativas al uso de la IA que tengan repercusiones significativas en la sociedad y el Estado deben apoyarse en una previsión científicamente validada e interdisciplinar de las consecuencias y riesgos socioeconómicos, así como en un examen de los posibles cambios en los valores sociales y el desarrollo cultural [17] [20].

2.2 Actitud responsable

En cada fase del ciclo de vida de los sistemas de IA, los actores deben adoptar una postura responsable respecto a los componentes que tienen un impacto en

la sociedad. Por ejemplo, la protección de la privacidad, el uso de la información personal de forma ética, segura y responsable, la evaluación del tipo, el alcance y la magnitud del daño potencial derivado, así como la elección y utilización de hardware y software auxiliares [20] [22].

2.3 Precaución

Cuando las acciones de un actor de IA tengan el potencial de tener efectos éticamente adversos sobre las personas y la sociedad, deberán tomarse medidas precautorias para evitar o minimizar la manifestación de tales efectos [19].

2.4 No hacer daño

No debe permitirse el uso de tecnologías, diseño, desarrollo, pruebas, implementación o el funcionamiento de IA con el objetivo de poner en peligro la vida humana, el medio ambiente o la salud o la propiedad de ciudadanos y entidades legales [20].

2.5 Identificación de IA en la comunicación con personas

Se insta a que se asegure de que los usuarios sepan que la interacción se realiza con una IA y no con una persona; y de que se informe cuando se afecten los derechos u otros aspectos importantes de la vida de quienes utilizan sistemas inteligentes [19].

2.6 Seguridad de los datos

Al utilizar sistemas de IA, los actores deben cumplir la legislación nacional pertinente en materia de datos personales y secretos protegidos. Los desarrolladores son responsables de salvaguardar y proteger los datos personales procesados por los sistemas de IA o por ellos mismos. Además, se espera que apliquen e incorporen técnicas avanzadas para impedir el acceso no autorizado a los datos personales por parte de terceros. Por otra parte, los actores de la IA deben garantizar la utilización de conjuntos de datos representativos y de alta calidad adquiridos legalmente de fuentes fiables [19].

2.7 Seguridad de la información

Mediante la aplicación de tecnologías de seguridad de la información adecuadas, como el empleo de mecanismos internos para salvaguardar los sistemas de IA de intervenciones no autorizadas y la notificación a los usuarios y desarrolladores de cualquier anomalía, los actores están obligados a garantizar la máxima protección

contra las vulnerabilidades en el funcionamiento de la IA. Asimismo, los usuarios deben conocer las políticas de seguridad de la información de estos sistemas.

2.8 Certificación voluntaria y cumplimiento del Código

Los actores de la IA tienen la opción de certificar voluntariamente que las tecnologías de IA creadas cumplen las normas establecidas en el Código ético de la IA [20] y en la legislación de la Federación Rusa [21].

2.9 Control de la automejora recursiva de los sistemas de IA

Rusia insta a los agentes de la IA a que colaboren para identificar y validar técnicas de desarrollo de IA general o fuerte, así como para prever los peligros potenciales que puedan presentar sistemas autónomos. El Estado debe supervisar la aplicación de las tecnologías de IA general [19]. Como ya se dijo anteriormente, la inteligencia artificial general (AGI) se refiere a la capacidad de un agente no humano para comprender o aprender cualquier tarea intelectual que puedan hacer los seres humanos. Se trata de la representación de capacidades cognitivas humanas generalizadas en software para que, ante una tarea desconocida, la máquina pueda encontrar una solución. La AGI sería capaz de entender el mundo tan bien como cualquier persona, en cambio la IA estrecha está diseñada para tareas específicas [23].

3. Responsables de las consecuencias del uso de IA

3.1 Supervisión

Dependiendo del objetivo de los sistemas, los actores de la IA deben proporcionar una supervisión humana exhaustiva en el grado y la forma que se requiera, por ejemplo, registrando las principales decisiones tomadas a lo largo del ciclo de vida de los sistemas o creando facilidades para su registro [19].

3.2 Responsabilidad

Los desarrolladores no deben otorgar a los sistemas de IA la capacidad de tomar decisiones moralmente responsables, ni culparlos por los resultados. Las implicaciones serán siempre responsabilidad de una persona, ya sea física o jurídica reconocida como sujeto de obligaciones de acuerdo con la legislación de la Federación Rusa [24].

4. Tecnologías de IA donde beneficien a las personas

4.1 Aplicación de los sistemas de acuerdo con su propósito

Los actores de la IA deben emplear los sistemas en el ámbito temático prescrito para las dificultades que se presenten y de acuerdo con el objetivo declarado.

4.2 Estimular el desarrollo de la IA

Los actores de la IA deben apoyar y recompensar la creación, la aplicación y el avance de una tecnología de IA ética y segura, teniendo en cuenta al mismo tiempo los intereses regionales y nacionales [19].

5. Dar mayor prioridad al desarrollo de la IA que a la competencia

5.1 Corrección de las comparaciones de la IA

Los actores de la IA deben emplear los datos más precisos y comparables sobre las capacidades en relación con una tarea y garantizar la uniformidad de la metodología de medición para mantener una competencia leal y una colaboración fructífera entre desarrolladores [19].

5.2 Desarrollo de competencias

Se anima a los actores de la IA a adherirse a las normas establecidas por la profesión, a mantener el nivel adecuado de compromiso para un trabajo seguro, eficaz y, en el marco de iniciativas y ámbitos académicos sobre ética, avanzar en la mejora de la competencia profesional de los involucrados en el campo de la IA [20]

5.3 Colaboración con los desarrolladores

En particular se insta a los desarrolladores a cooperar dentro de la comunidad, compartiendo entre sí información sobre vulnerabilidades importantes para erradicarlas. Debe crearse las condiciones para la formación de una “Escuela nacional de desarrollo de tecnologías de IA”. Ésta, según la iniciativa rusa, debería incluir repositorios nacionales de bibliotecas y modelos de redes de acceso público y de fácil acceso [25].

6. Transparencia sobre capacidades y riesgos de la IA

6.1 Credibilidad de la información

Se insta a los actores de la IA a que proporcionen a los usuarios información fiable sobre la tecnología y las mejores prácticas al utilizarla; del mismo modo, los riesgos y restricciones asociadas a su uso.

6.2 Sensibilización sobre la ética de la aplicación de la IA

Realizar esfuerzos para aumentar la confianza y la conciencia de responsabilidad. Esto debería implicar la aplicación moral a los sistemas y la redacción de artículos para revistas, planificación de seminarios y conferencias, tanto para el público en general como para los científicos, y la incorporación de normas y directrices éticas entre usuarios y operadores de los sistemas [19].

Hasta aquí los seis principios y sus aplicaciones específicas al desarrollo y uso de sistemas de inteligencia artificial en la Federación Rusa. Debe advertirse que el "Código de ética para la inteligencia artificial" fue elaborado por el Estado ruso y la comunidad científica, en estrecha colaboración de las principales "Empresas de la alianza en el ámbito de la IA" [Альянс в сфере искусственного интеллекта] [26]. Entre estas destacan: Sberbank, Yandex, MTS, VK, RDIF y Gazpromneft.

- Sberbank es una empresa de servicios bancarios y financieros de propiedad mayoritariamente estatal con sede en Moscú.
- Yandex es una multinacional tecnológica rusa que ofrece productos y servicios relacionados con Internet, incluido su motor de búsqueda y se le conoce como el "Google ruso" desde 1997.
- MTS (Mobile TeleSystems) es el mayor operador de telefonía móvil de Rusia y presta servicios a más de 80 millones de clientes; también ofrece una amplia gama de servicios digitales y posee diversos medios de comunicación.
- VK, antes conocida como VKontakte, es una red social rusa con sede en San Petersburgo. Cuenta con más de 100 millones de usuarios y ofrece funciones de comunicación, entretenimiento, negocios y noticias compartidas desde cualquier parte del mundo. Es la segunda mayor empresa de Internet en Rusia, después de Yandex.
- RDIF es el Fondo Ruso de Inversión Directa y es el fondo soberano de aquel país, creado en 2011 por el Gobierno para realizar inversiones en empresas

y atraer inversiones extranjeras en ámbitos sociales como la sanidad, el medio ambiente y otros mercados.

- Gazpromneft es una filial de Gazprom, una empresa dedicada al petróleo y gas que explora, desarrolla y produce crudo y gas natural. Es el tercer productor de petróleo de Rusia y ocupa también el tercer puesto en cuanto a producción de refinados del crudo.

Discusión

De acuerdo con la postura del Gobierno de la Federación Rusa, el desarrollo de tecnologías de IA debe dar prioridad a la protección de los derechos y libertades de las personas y los grupos. Esto implica adherirse a principios como la seguridad de la información, la no discriminación, la transparencia y el manejo responsable de los datos.

El uso de la IA debe guiarse por el principio de beneficiar a las personas y asumir la responsabilidad de las consecuencias de las aplicaciones de la IA. Por ello, es crucial garantizar que los ciudadanos estén plenamente informados de las oportunidades, los riesgos, así como de los éxitos y fracasos de la tecnología [27].

En Rusia las autoridades gubernamentales están de acuerdo con Occidente en que el uso generalizado de la IA sin una regulación ética puede conducir a la discriminación, el daño a los seres humanos, la pérdida de privacidad y control, errores algorítmicos y el uso de ésta para fines maleficentes. Estos puntos están en consonancia con la literatura especializada que hace énfasis en la necesidad de un desarrollo y despliegue responsables y éticos de la IA [28].

Más allá de las iniciativas externadas el 26 de octubre de 2021 durante el foro "Ética de la inteligencia artificial: el comienzo de la confianza" [18], el gobierno ruso tiene una estrategia nacional para el desarrollo de IA, aprobada en octubre de 2019 por el presidente Putin, denominado "Decreto 490 sobre el desarrollo de inteligencia artificial en la Federación Rusa" (Указ Президента Российской Федерации 490) [41] que tiene como objetivo convertir a Rusia en una potencia líder en IA para 2030 y utilizar la IA para proteger los intereses nacionales e implementar prioridades estratégicas, incluso en el ámbito de la defensa y la seguridad, más allá de los preceptos del Foro [15].

La estrategia, como aquí se ha descrito, aborda varias metas y objetivos para desarrollar capacidades de IA, tales como crear un entorno legal favorable, apoyar la investigación, la innovación, mejorar la educación, la capacitación digital, garantizar

la disponibilidad, la calidad de los datos, el desarrollo de infraestructura, promover la cooperación internacional y abordar cuestiones axiológicas y desafíos sociales.

También el gobierno ruso ha asignado roles y responsabilidades a diversas instituciones oficiales, como secretarías gubernamentales, agencias, empresas estatales, universidades y centros de investigación. Uno de los principales actores en la implementación de la estrategia es el "Centro Nacional de IA" [42], que se puso en marcha en septiembre de 2022 como una plataforma para coordinar proyectos de IA en diferentes sectores y dominios.

Sin embargo, pese a todos estos esfuerzos, Rusia enfrenta desafíos importantes para lograr sus metas en IA, como financiación limitada, fuga de cerebros, falta de confianza en actores del sector privado, así como la presión internacional por la guerra contra Ucrania.

República Popular (de) China

Estando China constantemente cuestionada desde Occidente sobre sus prácticas de desarrollo tecnológico y uso intensivo de la IA para vigilar a sus ciudadanos, el gobierno de la República Popular, a través de un Comité Estatal de Gobernanza, publicó en el mes de septiembre de 2021 un documento que establece normas éticas para el uso de la IA en aquel país titulado "Código de ética de la inteligencia artificial de nueva generación" (新一代人工智能伦理规范) [30]. "El código ético propone seis requisitos éticos básicos: mejorar el bienestar humano, promover la equidad y la justicia, proteger la privacidad y la seguridad, garantizar la controlabilidad y la fiabilidad, reforzar la responsabilidad y mejorar la alfabetización ética" [30, p. 1].

Se advierte que el propósito del código es "promover la equidad, la justicia, la armonía y la seguridad, y evitar problemas como los prejuicios, la discriminación, la privacidad y la fuga de información" [30, p. 1]. También se enumera que todo aquel que se dedique a la gestión, investigación, desarrollo, suministro y uso de IA debe observar los preceptos éticos enumerados en el *código. El suministro es definido como* producción, explotación y venta de productos y servicios de IA; mientras la gestión es "la planificación estratégica, la formulación y aplicación de políticas, reglamentos y normas técnicas, asignación de recursos, supervisión y revisión de la IA" [30, p. 1].

Los principios éticos que se proponen en el documento son seis características básicas: a) Promover el bienestar humano; b) promover la equidad y la justicia; c) proteger la privacidad y la seguridad; d) garantizar el control y la credibilidad; e) reforzar la responsabilidad; y f) fomentar la alfabetización ética.

El código se divide en normas de gestión, normas de I+D, normas de suministro y normas de uso.

Normas de gestión

1. Promover una gobernanza ágil

Para el gobierno chino es fundamental respetar las leyes del desarrollo de la IA, comprender sus posibilidades y limitaciones y mejorar constantemente los marcos y métodos de gobernanza. No alejarse de la realidad y tampoco centrarse únicamente en las ganancias y el éxito a corto plazo a la hora de tomar decisiones estratégicas, así como crear instituciones y asignar recursos suficientes [30, Art. 5].

2. Practicar activamente demostraciones

El código declara que se deben respetar las leyes, políticas y normas relativas a la IA, e integrar la ética en el proceso de gestión. Organizar a quienes practican y promueven la gobernanza tecnológica, sintetizar y compartir los conocimientos sobre IA, y abordar activamente las preocupaciones éticas planteadas por la sociedad [30, Art. 6].

3. Ejercicio correcto de la autoridad

Para el Estado chino se deben establecer condiciones y procesos normativos para el ejercicio del poder y definir con precisión las responsabilidades oficiales y los límites de autoridad de las actividades de gestión relacionadas con la IA. Respetar y garantizar las libertades, dignidad y seguridad de las personas. Impedir el uso indebido de la fuerza para violar las garantías individuales y los derechos de las empresas y otras entidades [30, Art. 7].

4. Fortalecer la prevención de riesgos

La IA debe desarrollarse y utilizarse de forma segura y responsable. Para lograr esto, los usuarios deben ser conscientes de los riesgos que conlleva el uso de la IA y deben asegurarse de tener un plan para hacer frente a los posibles problemas

que puedan surgir. Los investigadores, desarrolladores y responsables deben evaluar los riesgos asociados con el desarrollo de la IA, establecer un sistema de monitoreo y alertar a los usuarios a tiempo en caso de amenazas [30, Art. 8].

5. Promover la inclusión y la apertura

Para China se debe promover la tolerancia y la apertura a través de respetar los derechos, intereses y reivindicaciones de todas las partes involucradas; promover el uso de tecnologías de IA para abordar problemas reales de desarrollo económico y social; estimular la colaboración y la comunicación entre diversos ámbitos, regiones y fronteras y, por último, fomentar la creación de marcos y normas de gobernanza de la IA [30, Art. 9]. China está dividida en un total de 34 divisiones de la siguiente manera: 23 provincias; 4 municipalidades; 5 regiones autónomas y 2 regiones administrativas especiales.

Sobre normas de investigación y desarrollo

6. Fortalecer la conciencia de autodisciplina

Las Normas destacan la importancia de reforzar la autocontención en las actividades de investigación y desarrollo de la IA, así como de incorporar activamente la ética de la IA en cada etapa. En otras palabras, es fundamental construir una cultura de desarrollo responsable, en la que los investigadores y desarrolladores incluyan consideraciones éticas en todo el proceso.

Para lograr lo anterior, se recomienda que los investigadores y desarrolladores de IA realicen autoexámenes y refuercen la autogestión, esto significa que deben ser conscientes de las posibles implicaciones éticas de su trabajo. La IA tiene el potencial de tener un impacto significativo en la sociedad, por lo que es responsabilidad de quienes participan en su desarrollo asegurarse de que se haga de forma responsable y ética [30, Art. 10].

7. Mejorar la calidad de los datos

Cumplir todas las leyes, normas y estándares relacionados con recopilar, almacenar, utilizar, transmitir, suministrar y otras fases similares de procesamiento de los datos; mejorar su corrección, coherencia, puntualidad y conformidad normativa [30, Art. 11].

8. Mejorar la seguridad y la transparencia

Es importante asegurarse de que los sistemas de inteligencia artificial sean sólidos y resistentes a las interferencias. Para lograr esto, se deben mejorar diferentes aspectos del diseño, implementación y aplicación de los algoritmos utilizados [30, Art. 11].

La interferencia se refiere a un fenómeno en el que el rendimiento o el comportamiento de un modelo de aprendizaje automático se ve afectado negativamente por la presencia de información irrelevante o conflictiva durante el entrenamiento. La interferencia puede producirse de varias maneras y manifestarse como una disminución de la precisión, un aumento de las tasas de error o una degradación general del rendimiento del modelo.

9. Evitar la discriminación por prejuicios

Tener en cuenta la ética al desarrollar y utilizar la IA. El gobierno chino recomienda realizar investigaciones éticas más sólidas y tener en cuenta las diferentes perspectivas al recopilar datos y desarrollar algoritmos.

Es importante evitar cualquier tipo de sesgo en los datos o algoritmos, y trabajar hacia un sistema de IA que sea inclusivo, equitativo y no discriminatorio. En otras palabras, se trata de asegurarse de que la inteligencia artificial no discrimine a ciertas personas o grupos por sus características: raza, situación socioeconómica, credo, educación, sexo, entre otras [30, Art. 13].

10. Respetar las reglas del mercado

China recomienda garantizar el cumplimiento de las normas y reglamentos de acceso al mercado, las directrices de competencia y los protocolos de transacción, al tiempo que se promueve activamente un entorno de mercado saludable para el desarrollo de la IA. Evitar perturbar la competencia leal en el mercado impidiendo los monopolios de datos y plataformas. Prohibir cualquier infracción de los derechos de propiedad intelectual de otras entidades por cualquier medio [30, Art. 14]. Sin embargo, se debe señalar que, en el pasado, el país ha sido acusado de violar los derechos de propiedad intelectual industrial.

Los siguientes artículos pueden ser reiterativos, toda vez que refieren al control de la calidad [Art. 15]; protección de los derechos e intereses de los usuarios [Art. 16]; tener una buena respuesta a las emergencias [Art. 17]; promover la buena voluntad en todos los procesos [Art. 18]; evitar el mal uso y abuso de la IA [Art. 19], así como una buena retroalimentación entre usuarios y desarrolladores [Art. 20].

Como puede observarse, China en un breve código de 20 artículos aborda prácticamente todos los aspectos concernientes a la gobernanza, uso y desarrollo de la inteligencia artificial de manera ética y responsable.

Discusión

China se ha fijado el objetivo de convertirse en líder mundial en el desarrollo de la inteligencia artificial para 2030 [32]. En 2017, el Consejo de Estado Chino ya había publicado el "Plan de desarrollo de la inteligencia artificial de nueva generación" (新一代人工智能发展规划的通知) [33], que trazó una hoja de ruta para el desarrollo de la IA en aquel país. En 2021 actualizó su normatividad a través del "Código de ética de la inteligencia artificial de nueva generación" (新一代人工智能伦理规范) [30]. Tanto el Plan como el Código identifican la IA como una tecnología estratégica para el desarrollo económico y social de ese país, y esbozan una serie de ambiciosos objetivos para su industria de la IA, entre ellos:

- Convertirse en el principal centro de innovación de IA del mundo para 2030.
- Construir una cadena industrial de IA completa de alta competitividad.
- Lograr grandes avances en las teorías, algoritmos e investigación básica de la IA.
- Promover la integración de la IA con otras industrias para crear nuevas formas de fabricación, servicios y productos inteligentes.
- Establecer un sistema integral de gobernanza de la IA para garantizar su desarrollo seguro y responsable [30].

Para alcanzar estos objetivos, el gobierno chino ha realizado importantes inversiones en investigación, promovido la comercialización de tecnologías y fomentado la cooperación internacional en el desarrollo. Lo anterior se considera un motor clave del crecimiento económico y el progreso tecnológico de China, por lo que el gobierno ha identificado la IA como una prioridad en sus planes a largo plazo.

La estrategia también establece fomentar ecosistemas de innovación, construir infraestructura inteligente, promover aplicaciones industriales, fortalecer el cultivo de talento, mejorar los sistemas de gobernanza y expandir la cooperación internacional.

Además, el gobierno de Xi Jingpin ha asignado roles y responsabilidades a diversas instituciones estatales, como secretarías, agencias, gobiernos locales, empresas, universidades e institutos de investigación. Uno de los principales actores en la

implementación de la política es el Ministerio de Ciencia y Tecnología [34], que supervisa diversos proyectos y plataformas de IA a nivel nacional.

Sin embargo, algunos expertos han señalado que China tiene desafíos importantes para lograr sus ambiciones, como lo son algunas preocupaciones éticas y sociales, especialmente problemas de privacidad y derechos humanos, cuellos de botella técnicos, como la calidad y seguridad de los datos; incertidumbres regulatorias, como violación de estándares y normas. China enfrenta además guerras comerciales especialmente con los Estados Unidos [35] y conflictos geopolíticos con otros países, como sus vecinos India, Japón y Vietnam, así como con su propio territorio, Taiwán [36].

Conclusiones parciales

No hay un país que tenga un liderazgo indiscutible en inteligencia artificial, ya que es un campo que evoluciona y se desarrolla con rapidez. Algunos países están invirtiendo mucho en investigación de la IA, y los avances dependen de diversos factores, como las políticas gubernamentales, la financiación, el talento y las infraestructuras.

Los códigos éticos sobre IA que han redactado Estados Unidos, Rusia y China siguen evolucionando y están sujetos a interpretación. La comunidad internacional debate sobre los mejores enfoques de la ética de la IA, y es probable que estos códigos cambien en respuesta a los avances tecnológicos y los nuevos escenarios. Cada uno de estos países le ha dado su propio enfoque de acuerdo con sus intereses, sus posiciones geopolíticas e incluso su cosmovisión.

Estados Unidos tal vez sea la nación que más ha avanzado en este sentido, con la "Ley de iniciativa nacional de inteligencia artificial" (NAIIA - Act of 2020, Div. E, Sec. 5001) [1], que establece un programa coordinado en todo el gobierno federal para acelerar la investigación en IA para la aplicación sistemática, prosperidad económica y la seguridad nacional de aquel país.

China, por su parte, desea cumplir su meta de ser el líder mundial en materia de inteligencia artificial para 2030 y para ello ha implementado el "Plan de desarrollo de inteligencia artificial de nueva generación" de 2017 y el "Código de ética de la inteligencia artificial de nueva generación" de 2021. Cabe resaltar que China asumió una agresiva estrategia de expansión de desarrollo, investigación e implementación de IA, tanto en el sector público como privado. Por ejemplo, el gobierno

de Xi Jinping está apoyando decididamente a las empresas emergentes en diversos rubros que utilizan la IA; por su parte, el gobierno mismo utiliza diversos sistemas, como el reconocimiento facial en vigilancia y el Sistema de crédito social, del que ya se habló anteriormente.

Rusia también ha hecho esfuerzos significativos por no quedar atrás en esta guerra fría de la IA, pues las consecuencias de llevar el liderato son muchas. Tanto en lo comercial, industrial, defensa y tecnología la IA significa una ventaja competitiva y comparativa que ninguno de los tres quiere ceder.

Otros países como Canadá [37], el Reino Unido [38], Alemania [39] y Japón [40] también están invirtiendo muchos recursos en investigación y desarrollo de IA y han realizado importantes contribuciones para la construcción de un marco ético normativo. En México, el 30 de marzo de 2023 se presentó en la Cámara de Diputados una "Iniciativa con proyecto de decreto por el que se expide la Ley para la regulación ética de la inteligencia artificial y la robótica para los Estados Unidos Mexicanos" [43], la cual propone la creación de un "Consejo mexicano de ética para la inteligencia artificial"; sin embargo, la iniciativa, aún no ha obtenido la atención de las Comisiones Legislativas respectivas.

A grandes rasgos, puede decirse que que carrera de la IA ha comenzado y habrá mucha tensión entre las normativas nacionales y el "campo de batalla" tecnológico, industrial, comercial e incluso militar de aplicaciones presentes y futuras. En última instancia, el liderazgo en IA dependerá de la continuidad de la inversión, la investigación y el desarrollo, y de la capacidad para aprovechar la tecnología para beneficio económico y social de cada país.

En conclusión, las normativas y políticas de Estado para una ética de la IA de Estados Unidos, China y Rusia reflejan el creciente reconocimiento de la necesidad de regular y gobernar el desarrollo y despliegue de esta poderosa tecnología. Aunque estas propuestas varían en cuanto a su enfoque, todas reconocen los beneficios y riesgos potenciales de la IA y la necesidad de que su desarrollo se rija por consideraciones acordes, por ejemplo, a los Tratados de Ginebra que establecen normas jurídicas para el trato humanitario en la guerra.

A medida que estas tres superpotencias compiten y cooperan en un mundo cada vez más interconectado, es esencial que den prioridad a la colaboración y al diálogo sobre la ética de la IA para garantizar que esta tecnología se desarrolle y despliegue de forma coherente con los valores humanos y el bien común, no solo estratégico y geopolítico. Solo a través de un compromiso compartido de una IA

responsable y ética se puede esperar evitar una escalada bélica y, en su lugar, construir un futuro en el que la IA beneficie a la especie humana.

Referencias

[1] USA, "National AI Initiative Act of 2020 (DIVISION E, SEC. 5001)," AI.gov. Acceso jun. 2023. [En línea] Disponible: https://www.ai.gov/

[2] USA, "The National Artificial Intelligence Initiative Office (NAIIO)," AI.gov. Acceso jun. 2023. [En línea] Disponible: https://www.ai.gov/naiio/#ABOUT-NAIIO

[3] W. House, "The Biden Administration Launches the National Artificial Intelligence Research Resource Task Force," Whitehouse.gov. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3y

[4] F. Bignami, "Artificial Intelligence Accountability of Public Administration," *Am. J. Comp. L.*, vol. 70, no. Supplement_1, pp. i312-i346, 2022, doi: 10.1093/ajcl/avac012

[5] E. Brown, "Copy of Policy 101: Artificial Intelligence / Machine Learning_ATPH," Aspen Tech Policy Hub, United States of America, 2022. Acceso jun. 2023. [Online]. Disponible: https://policycommons.net/artifacts/3360862/copy-of-policy-101/4159555/ https://bsu.buap.mx/b3y https://bsu.buap.mx/b3y

[6] U.S. Department of State, "Artificial Intelligence (AI)," State.gov. Acceso jun. 2023. [En línea] Disponible: https://www.state.gov/artificial-intelligence/

[7] S. Hristova, "Proto-Algorithmic War. How the Iraq War became a laboratory for algorithmic logics." Social and Cultural Studies of Robots and AI (SOCUSRA). Londres: Pelgrave, 2022, doi: 10.1007/978-3-031-04219-5

[8] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, y L. Floridi, "Artificial Intelligence and the 'Good Society': the US, EU, and UK approach," *Sci. and Eng. Ethics*, vol. 24, no. 2, pp. 505-528, Feb. 2018, doi: 10.1007/s11948-017-9901-7

[9] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo and L. Floridi, "The ethics of algorithms: key problems and solutions," *AI & Soc.*, vol. 37, no. 1, pp. 215-230, Feb. 2022. doi: 10.1007/s00146-021-01154-8

[10] L. Floridi et al., "An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed. Cham: Springer International Publishing, 2021, pp. 19-39.

[11] C. Stadlmann y A. Zehetner, "Human Intelligence Versus Artificial Intelligence: A Comparison of Traditional and AI-Based Methods for Prospect Generation," en *Marketing and Smart Technologies*, NY: Springer, 2021, pp. 11-22, doi: 10.1007/978-981-33-4183-8_2

[12] V. Yazdanpanah, E.H. Gerding, S. Stein, M. Dastani, C.M. Jonker, T.J. Norman, y S.D. Ramchurn, "Reasoning about responsibility in autonomous systems: challenges and opportunities," *AI & Soc.* dic., 2022, doi: 1007/s00146-022-01607-8

[13] D. Acemoglu, D. Autor, J. Hazell y P. Restrepo, “Artificial Intelligence and Jobs: Evidence from Online Vacancies,” J. Labor Econ., vol. 40, dic. 2022. https://bsu.buap.mx/ciR

[14] BBC, “Chernihiv: Are these Russia’s weapons of war?” BBC.com. Acceso jun. 2023. [En línea] Disponible: https://www.bbc.com/news/world-europe-61036880

[15] DeutscheWelle, “Russia steps up use of kamikaze drones in Ukraine,” DW.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3z

[16] AP, “Putin: Leader in artificial intelligence will rule world,” AP.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3A.

[17] Rusia, “Una estrategia nacional de IA,” [Кодекс этики в сфере искусственного интеллекта], 2021. [En línea] Disponible en: https://bsu.buap.mx/ch0

[18] Rusia, “I Foro Internacional ‘Ética de la inteligencia artificial: el comienzo de la confianza’,” [I международный форум “Этика искусственного интеллекта: начало доверия”]. A-ai.ru. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3B

[19] Rusia, “AI Ethics Code,” [Кодекс этики в сфере ИИ], Tass.com. Acceso jun. 2023. [En línea] Disponible: https://ethics.a-ai.ru/

[20] Rusia, “Código Ético de la Inteligencia Artificial” (Кодекс этики в сфере искусственного интеллекта). Ai.gov.ru. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ch0

[21] Rusia, “Concepto Regulatorio” (Концепция нормативного регулирования) 2022. [En línea]. Disponible en: https://ai.gov.ru/regulation/kontseptsiya-normativnogo-regulirovaniya/ y Концепция развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 года. Acceso jun. 2023. [En línea] Disponible en: http://government.ru/docs/all/129505/

[22] Y. Kharitonova, V. Savina, y F. Pagnini, “Civil Liability in The Development And Application Of Artificial Intelligence And Robotic Systems: Basic Approaches,” *Bulletin of the Univ. of Perm. Legal Sci*., vol. 58, pp. 683-708, 2022, https://bsu.buap.mx/ci5

[23] C. Turner and S. Schneider, “Could You Merge with AI?: Reflections on the Singularity and Radical Brain Enhancement,” in *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford University Press, 2020, pp. 307-324.

[24] Rusia, “Decretos y resoluciones,” [Указы и постановления]. AI.gov.ru. Acceso jun. 2023. [En línea] Disponible: https://ai.gov.ru/strategy/uip/.

[25] Rusia, “Estandarización,” [Стандартизация]. AI.gov.ru. Acceso jun. 2023. [En línea] Disponible https://ai.gov.ru/regulation/standardization/

[26] Tass, “First code of ethics of artificial intelligence signed in Russia,” Tass.com. Acceso jun. 2023. [En línea] Disponible: https://tass.com/economy/1354187

[27] B. Ammanath, Trustworthy AI: A Business Guide for Navigating Trust and Ethics in AI. NJ, USA: Wiley, 2022.

[28] L. Floridi, “Establishing the Rules for Building Trustworthy AI,” in *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed. Cham: Springer International Publishing, 2021, pp. 41-45.

[29] European Commission, “Cross-cutting Thematic Event, The impact of the Russian war of aggression against Ukraine on P/CVE,” Commission.europa.eu. Acceso jun. 2023. [En línea] Disponible https://bsu.buap.mx/b3F

[30] China, “Lanzamiento del ‘Código de Ética de Inteligencia Artificial de Nueva Generación’,” [新一代人工智能伦理规范] (En chino) 2021. Most.gov. Acceso jun. 2023. [En línea] Disponible: https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html [enlace permanente: https://bsu.buap.mx/coi]

[31] S. Huai, “Un artículo explica claramente cómo mejorar la calidad del big data,” Cloud.com. Acceso jun. 2023. [En línea] Disponible: https://cloud.tencent.com/developer/news/609423

[32] S. O’Meara, “Will China lead the world in AI by 2030?” Nature.com. Acceso jun. 2023. [En línea] Disponible: https://www.nature.com/articles/d41586-019-02360-7

[33] China, “Aviso del Consejo de Estado sobre la impresión y distribución del Plan de desarrollo para una nueva generación de inteligencia artificial,” 2017. Gov.cn. Acceso jun. 2023. [En línea] Disponible: https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

[34] China, “Ministerio de Ciencia y Tecnología de la República Popular China,” Gov.cn. Acceso jun. 2023. [En línea] Disponible: https://www.most.gov.cn/index.html

[35] H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang, y L. Floridi, “The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation,” *AI & Soc.* Vol. 36, pp. 59–77, mar. 2021, doi: 10.1007/s00146-020-00992-2

[36] L. Silver, C. Huang, y L. Clancy, “Negative Views of China Tied to Critical Views of Its Policies on Human Rights,” Pewresearch.org. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b3I

[37] T. Lepage-Richer and F. McKelvey, “States of computing: On government organization and artificial intelligence in Canada,” *Big Data & Soc.*, vol. 9, no. 2, p. 20539517221123304, 2022, https://bsu.buap.mx/ci7

[38] J. Reis, P.E. Santo, y N. Melão, “Artificial intelligence in government services: A systematic literature review,” in NKIST: Volume 1, 2019, pp. 241-252, doi: 10.1007/978-3-030-16181-1_2

[39] M. Vergeer, “Artificial intelligence in the Dutch press: An analysis of topics and trends,” *Comm. Stud.*, vol. 71, no. 3, pp. 373-392, 2020, doi: 10.1080/10510974.2020.1733038

[40] C. Narvaez Rojas, et al., “Society 5.0: A Japanese concept for a superintelligent society,” Sustainability, vol. 13, no. 12, p. 6567, 2021.

[41] Rusia. “Decreto del Presidente de la Federación de Rusia sobre el desarrollo de la inteligencia artificial No. 490” (Указ Президента Российской Федерации 490). Kremlin.ru. Acceso jun. 2023. [En línea] Disponible: http://www.kremlin.ru/acts/bank/44731

[42] Pockom, “Centro Nacional de Inteligencia Artificial inaugurado en Rusia” (В России начал работу Национальный центр искусственного интеллекта). 2022. En línea: https://dzen.ru/a/YxsKpP7f0GsfmeJN

[43] I. Loyola. “Iniciativa con proyecto de decreto por el que se expide la Ley para la Regulación Ética de la Inteligencia Artificial para los Estados Unidos Mexicanos”. Gobernacion.gob.mx. 30 de marzo de 2023. Acceso jun. 2023. [Disponible en línea]: https://bsu.buap.mx/ch6

[44] T. Sterling y S. van den Berg, “Ukraine war shows urgency of military AI, Palantir CEO says,” Reuters.com. 15 feb. 2023. Acceso jun. 2023. [Disponible en línea]: https://bsu.buap.mx/clp

MENOS, ES MÁS: RECONSTRUIR UNA ÉTICA CLÁSICA NORMATIVA PARA UN FUTURO RESPONSABLE DE LA INTELIGENCIA ARTIFICIAL

Introducción

La repetición y la superposición innecesaria de principios éticos similares para el desarrollo de una inteligencia artificial responsable no solo entran en conflicto, sino que esta confusión y ambigüedad pueden llegar, incluso, a resultar peligrosas si los postulados son un mero "lavado de cara" y las verdaderas intenciones se esconden detrás de intereses mezquinos. Esto aplica tanto a particulares, a empresas, como a gobiernos. El proceso de establecer leyes, normas, estándares y mejores prácticas para asegurar que la IA sea benéfica para toda la sociedad es un llamado urgente para un "mejor futuro de la humanidad". De todo lo anterior, surge este intento por traducir en solo seis principios fundamentales el cúmulo de literatura que hasta ahora se tiene y, posteriormente, defender tres máximas que podrían sintetizar no solo los principios esbozados en este libro, sino los marcos normativos vigentes con aspiraciones universalistas: responsabilidad, transparencia, seguridad, no discriminación, privacidad y sostenibilidad; y como máximas clásicas de la ética: honradez, intencionalidad y conciencia moral.

Los primeros esfuerzos

Isaac Asimov propuso las Leyes de la robótica en una serie de cuentos cortos que escribió en la década de 1940. Los tres primeros cuentos que incluyen las leyes se titulan: Robbie; Runaround; y Reason [1]. En estas obras, reunidas en el libro *I, Robot* (Yo, Robot) de 1950, Asimov propuso tres leyes fundamentales que deben seguirse al desarrollar sistemas robotizados:

- Primera Ley: Un robot no puede dañar a un ser humano, o por inacción permitir que un ser humano sufra daño.

- Segunda Ley: Un robot debe cumplir las órdenes dadas por los seres humanos, a excepción de aquellas que entren en conflicto con la primera ley.
- Tercera Ley: Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o la segunda ley.

Las Leyes de la robótica se citan en numerosos artículos científicos y obras literarias de la ciencia ficción [2]. Hoy, han tomado especial relevancia en el debate mundial sobre la IA. Aunque no son propiamente leyes, en sentido jurídico que deban seguirse sobre pena de sanción, representan una reflexión ética sobre cómo debe utilizarse, controlarse y autocontrolarse los sistemas autónomos o semiautónomos.

Por otra parte, el denominado "Verano de Dartmouth de 1955 sobre inteligencia artificial" fue una Conferencia que tuvo lugar en el Dartmouth College en Hanover, New Hampshire, Estados Unidos [3]. Fue organizada por el matemático John McCarthy, quien es conocido como uno de los padres de la inteligencia artificial. La reunión fue un hito importante en el desarrollo de la IA, ya que fue el evento en el que se acuñó el término "inteligencia artificial" y se discutió en profundidad cómo crear computadoras que pudieran pensar y aprender del mismo modo en que lo hacen los seres humanos y algunas implicaciones para la ética [3].

Durante la Conferencia, McCarthy y otros participantes discutieron una variedad de temas relacionados, como el aprendizaje automático, la percepción y el lenguaje natural. El evento tuvo un gran impacto y sentó las bases para el desarrollo de muchos de los avances que se han realizado desde entonces.

En realidad, es difícil determinar quiénes fueron los primeros en discutir la ética de la IA, ya que el debate ha sido una preocupación constante a lo largo de su historia. Desde el momento en que se empezó a desarrollar la IA, se han planteado preguntas fundamentales sobre cómo deben utilizarse estos sistemas y qué consecuencias pueden tener para la sociedad y para los individuos.

Hoy, la inteligencia artificial es una tecnología que está transformando rápidamente la forma en que se vive y trabaja. A medida que la IA se vuelve más prominente en el mundo ontológico y digital, es importante que los creadores de sistemas asuman la responsabilidad de sus creaciones y se aseguren de que estos sistemas se utilicen de manera ética y responsable.

Esto incluye hacer sistemas explicables y transparentes en su funcionamiento, toma de decisiones, seguridad y protección contra el uso indebido. También se espera que no sean discriminatorios contra ningún individuo o grupo de personas y que respeten la privacidad de los usuarios. La brecha digital se está haciendo cada vez más grande entre quienes tienen y no tienen acceso a las nuevas herramientas.

Es por todo lo anterior por lo que se trata de traducir estas aspiraciones en los siguientes principios:

1. Responsabilidad

Es importante que los creadores de sistemas de IA asuman la responsabilidad de sus algoritmos y que se utilicen de manera ética. Debe haber claridad en quién es responsable de la IA y las decisiones que esta tome, y se deben establecer mecanismos para asegurar que cumpla con los estándares éticos y legales aplicables [4] [5].

La responsabilidad en programación se refiere a la necesidad de considerar cuidadosamente las consecuencias de crear y utilizar software. Esto incluye tanto la responsabilidad individual del programador de considerar el impacto de su trabajo como la responsabilidad de la industria de la programación de abordar cuestiones éticas, de impacto social y ambiental en el desarrollo y aplicación de IA, y crear sus propios instrumentos de regulación y respetarlos mientras no exista un marco normativo jurídicamente vinculante al respecto [6] [7] [8].

La cadena de responsabilidad en la creación y uso de IA incluye diferentes actores que tienen participación específica en el proceso [9]. Algunos sujetos de responsabilidad deben ser:

Los programadores, quienes son los responsables de crear el software y asegurarse de que cumpla con los estándares éticos y técnicos aplicables.

Los diseñadores de experiencias de usuario quienes son responsables de crear interfaces intuitivas, fáciles de manejar y considerar el impacto que su software pueda tener.

Los directores de proyecto quienes son responsables de gestionar y administrar el proyecto, así como asegurarse de que cumpla estándares éticos.

La empresa de software es responsable de crear, distribuir y comercializar sus productos [10].

Toda vez que la corporación y sus accionistas son los beneficiarios directos del desarrollo de los sistemas, deben asumir un compromiso ético y de representación legal. Se debe terminar con la idea errónea de que lo más importante es el precio de una acción, sin importar los mecanismos a través de los cuales se incrementa el precio de ese pagaré. En muchos casos el precio especulativamente aumenta por medidas draconianas que violan los derechos de los trabajadores, despidos injustificados, ahorros por encima del bienestar y la seguridad, buyback que es la autocompra de acciones para crear demanda, entre otras estrategias poco éticas [11].

Los usuarios también deben ser responsables de utilizar el software con fines benéficos, e informar sobre cualquier anomalía que detecten, tanto a la propia empresa como a las autoridades competentes. Por lo anterior, también deben existir entes reguladores como responsables de establecer y hacer cumplir las leyes y normas aplicables [12].

La importancia de reconocer que cada persona y puesto de trabajo implicado en la cadena de desarrollo y utilización de software tiene funciones y deberes distintos, obliga a construir un marco normativo general, pero a la vez específico para considerar todas las posibles circunstancias en las que podría estar involucrado un aspecto ético y, sobre todo, un punto de quiebre de decisión moral [13] [14].

Por ejemplo, los desarrolladores, los encargados de las pruebas de calidad, los gestores de proyectos y los usuarios finales pueden tener diferentes obligaciones y expectativas en relación con la creación y uso de la IA. Es esencial, por tanto, que todos comprendan sus responsabilidades y que las cumplan adecuadamente.

La norma de calidad estándar “IEEE 7000™-2021, proceso de modelo estándar IEEE para abordar preocupaciones éticas durante el diseño del sistema” [15], creada por el Instituto de ingenieros en electricidad y electrónica tiene como objetivo integrar los valores humanos y sociales en el diseño de sistemas tradicionales a través de procesos que permiten a los programadores traducir las consideraciones éticas y valores de las partes interesadas en requisitos del sistema y prácticas de diseño específicos. Este enfoque aborda las obligaciones regulatorias éticas en el diseño de sistemas inteligentes autónomos de manera sistemática y rastreable.

2. Transparencia

Los sistemas de IA deben ser transparentes y explicables, es decir, deben ser capaces de poder deconstruir el proceso de toma de decisiones. Esto es importante para evitar las cajas negras que escapan de la explicabilidad de sus propios creadores. Que una IA sea transparente significa que se puede visualizar su funcionamiento interno y la forma en que toma decisiones de manera comprensible para cualquiera con un mínimo de conocimientos. Algunos algoritmos, especialmente de aprendizaje automático y de decisiones estocásticas, esto es, tomadas por azar, quedan fuera incluso de la competencia y comprensión de quien los hizo [16] [17].

La transparencia es importante por varias razones: ayuda a los usuarios a entender cómo funciona el software y cómo pueden utilizarlo de manera adecuada; ayuda a detectar y solucionar problemas o errores internos; genera confianza en el uso del software, especialmente cuando se utiliza para tomar decisiones importantes o en contextos críticos; garantiza la cadena de responsabilidad y, por último, hace visibles las alteraciones mal intencionadas del sistema original [18] [19] [20] [21].

La transparencia en el software no es siempre una característica que se pueda medir de manera objetiva, es un concepto subjetivo que puede variar dependiendo del contexto y de las expectativas de los usuarios. Sin embargo, se debe alcanzar un justo medio entre la transparencia y la seguridad del algoritmo [18].

El hecho de que el software sea de código abierto también abre la posibilidad para el desarrollo de versiones mal intencionadas o usos indebidos. Toda tecnología, finalmente, puede tener un doble propósito [22].

2.1 Algoritmos de caja negra

Los algoritmos de caja negra o black boxes, como ya se explicó, son aquellos cuyo funcionamiento interno es desconocido, no puede ser explicado o no es comprensible. Esto significa que el algoritmo toma sus decisiones y procesa la información de manera oculta, incluso para su propio desarrollador, no necesariamente por mala fe, sino porque muchas de estas decisiones son tomadas a gran velocidad [23] [24]. Las redes neuronales, los algoritmos de aprendizaje automático supervisado, como las máquinas de vectores de soporte (SVM) [25], son ejemplos de estos algoritmos.

La relación entrada-salida de datos de un algoritmo de caja negra puede observarse, pero el funcionamiento interno está oculto. Por eso puede ser difícil

averiguar quién toma qué decisiones y por qué. En los casos en que las decisiones de los algoritmos tienen un efecto de gran alcance sobre las personas o la comunidad puede ser un problema grave no saber quién o cómo se tomó la decisión. Por ello, las decisiones cruciales deben quedar siempre en manos de personas [4].

Los algoritmos de caja negra son utilizados en diversas áreas, como la automatización de procesos, la toma de decisiones en sistemas de recomendación, asignación de etiquetas, clasificación y asignación de valores, así como en la detección de posibles fraudes en el sistema bancario, entre muchas otras aplicaciones. Aunque estos algoritmos pueden ser efectivos en algunas circunstancias, también pueden ser problemáticos debido a la falta de transparencia y, por ende, a la dificultad para identificar y corregir las fallas cuando se presentan [24].

Las redes neuronales en IA son un tipo de modelo computacional que imita el funcionamiento del cerebro humano. Es un sistema en el que se conectan entre sí nodos o neuronas artificiales y se activan al azar según los datos de entrada [26]. Por su parte, las SVM es un tipo de algoritmo de aprendizaje profundo que realiza aprendizaje supervisado para la clasificación o regresión de grupos de datos y pueden utilizarse para diversas tareas, como la clasificación de imágenes, la clasificación de textos, la identificación de escritura a mano, la detección de spam, el análisis genético, la identificación de rostros y la localización de cambios en los códigos [27].

Por lo anterior, la transparencia en el funcionamiento de los algoritmos es importante para garantizar la confianza, responsabilidad en su uso y evitar errores. Las redes neuronales pueden considerarse algoritmos de caja negra en el sentido de que el su funcionamiento interno no suele ser fácilmente interpretable por el ser humano. En otras palabras, puede ser difícil entender cómo una red neuronal llega a una salida o predicción concreta, dada una entrada específica. Esto se debe a que las redes neuronales suelen tener muchas capas de nodos interconectados, y cada nodo realiza una operación matemática compleja sobre los datos de entrada, lo que hace difícil seguir la secuencia de operaciones que conduce a una salida específica. Además, los pesos y sesgos de cada nodo se aprenden mediante un proceso de entrenamiento que puede ser complejo y muy poco intuitivo [32].

Sin embargo, se debe señalar que los investigadores han desarrollado técnicas para intentar comprender el funcionamiento de las redes neuronales, como la visualización de las activaciones de los nodos individuales, el uso del análisis de importancia de características para identificar cuáles son las más influyentes y el entrenamiento de modelos más pequeños e interpretables para imitar el comportamiento

de redes neuronales más grandes [58]. Aunque las redes neuronales pueden considerarse algoritmos de caja negra, se están haciendo esfuerzos por aumentar la comprensión de su funcionamiento interno, pero falta mucho por hacer.

Existen casos famosos de errores algorítmicos: la caída de los mercados en el año 2010 por un algoritmo que ejecutó órdenes especulativas de venta masivas de forma erráticas [28]; el accidente del Mariner 1, una sonda de la NASA que iba a sobrevolar Venus y que tuvo que ser derribada porque el algoritmo de navegación falló [31]. Mucho más grave por las pérdidas humanas fue la caída de los vuelos 610 de Lion Air a finales de 2018 [29] y 302 de Ethiopian Airlines en 2019 [30], ambos Boing 737 Max por culpa de un error de cálculo del software, por ambos accidentes 346 personas perdieron la vida. Finalmente, se debe señalar que entre 2016 y 2023, la Administración nacional de seguridad del tráfico en carretera de los Estados Unidos inició 41 investigaciones especiales de accidentes automovilísticos sospechosos de estar relacionados con el uso del sistema Autopilot de Tesla, en el que 19 personas perdieron la vida [60].

2.2 Equilibrio entre transparencia y seguridad

Una aspiración genuina debe ser encontrar el equilibrio entre la transparencia del software y su seguridad, porque ambos son aspectos fundamentales de su desarrollo. Algunas recomendaciones son:

- Proporcionar la información necesaria para que los usuarios y otros interesados conozcan el funcionamiento del algoritmo y comprendan sus decisiones, sin revelar información confidencial o sensible que pueda comprometer la seguridad [34].
- Establecer medidas adecuadas para proteger la privacidad y la seguridad de los usuarios para garantizar que el software no sea utilizado de manera maliciosa, o coseche datos que puedan ser utilizados para otros fines.
- Establecer mecanismos de revisión y evaluación independientes para asegurar que el sistema cumpla con los estándares éticos y de seguridad aplicables.
- Trabajar siempre con expertos en ética y en seguridad para identificar y abordar los conflictos que puedan surgir entre uno y otro campo.

Observando estos aspectos es posible conciliar el problema de una toma de decisiones rastreable y, al mismo tiempo, segura. Los algoritmos tienen responsabilidad limitada o nula, son sus desarrolladores quienes tienen que asumir un papel visible en este proceso [35].

3. Seguridad

Cuando se dice que los sistemas de IA deben ser seguros contra el uso indebido o la manipulación, se refiere a que deben protegerse diversos aspectos, por ejemplo:

La privacidad de los usuarios asegurándose de que la IA no haga uso indebido de la información personal; evitar que los sistemas causen daño o pongan en peligro a otros; y la integridad es que el software sea utilizado de manera honesta y que no sea fácilmente manipulado [18].

El papel fundamental de la seguridad y la protección en el desarrollo y despliegue de los sistemas de IA supone que, a medida que la tecnología se hace más avanzada y generalizada, se garantice que se utiliza de forma responsable y ética. Se deben aplicar las medidas y las garantías adecuadas que impidan el uso indebido o manipulación de los sistemas a través de barreras de fuego (firewall) difíciles de romper. Los firewalls han sido durante más de veinte años la primera fila de contención de ataques cibernéticos.

Sin embargo, entre las amenazas más peligrosas están la llamada ingeniería social, que depende más del error humano que de los fallos tecnológicos. Los ciberdelincuentes pueden persuadir a las personas para darles acceso a sistemas o redes sin autorización u obtener información sensible. Los hackers maliciosos engañan a sus víctimas haciéndose pasar por otra persona y aprovechándose de su curiosidad, miedo o persuasión obtienen, por ejemplo, contraseñas y otras credenciales. Los ataques de phishing, la suplantación de identidad, el robo de accesos y otros tipos de fraude en línea pueden llevarse a cabo mediante ingeniería social apoyada por inteligencia artificial [36].

El malware (o programas malignos), incluye numerosos peligros que van desde virus y gusanos hasta troyanos bancarios, adware (publicidad intrusiva), spyware (software espía) y ransomware (secuestro de datos) [37].

El objetivo de un ataque de denegación de servicio distribuido (DDoS) es interrumpir el tráfico normal de un servidor, servicio o red enviando un gran número de peticiones desde muchos dispositivos. El ataque satura el sistema e impide que los usuarios autorizados lo utilicen. Los ataques DDoS pueden ser llevados a cabo por personas o grupos con IA y pueden perjudicar gravemente a las empresas, gobiernos y organizaciones [38].

Por todo lo anterior, la ciberseguridad es el conjunto de prácticas e instrumentos utilizados para garantizar que los servicios, los datos generados y procesados por ordenadores, servidores, dispositivos móviles, redes y otros sistemas electrónicos funcionen correctamente y no puedan ser vulnerados [39].

Por tanto, las medidas de seguridad y protección deben ser parte integrante del proceso de desarrollo y uso de la IA. Se habla aquí del cifrado de datos, controles de acceso, auditoría y otras técnicas para evitar el acceso o uso no autorizados.

4. No discriminación

Un algoritmo puede ser discriminatorio si toma decisiones o proporciona resultados que favorecen a ciertos grupos en detrimento de otros basándose en características como la raza, género, orientación sexual, edad, discapacidad, datos demográficos, condición económica, ubicación, entre otros [40].

Esto puede ocurrir por varias razones: si el algoritmo es entrenado con datos que reflejan una discriminación previa, que muestran una mayor probabilidad de que ciertos grupos de personas sean rechazados para un determinado trabajo o servicio, el algoritmo podría reproducir esa discriminación al tomar decisiones. Por ejemplo, si el algoritmo tiene en cuenta ciertas variables para evaluar el riesgo de crédito como los códigos postales de los domicilios de las personas, género, edad u otros datos subjetivos, como preferencias de compra, entonces estará sesgado [41].

También, el algoritmo podría estar diseñado para discriminar a grupos de personas, por ejemplo, latinos, afrodescendientes o musulmanes, como suele suceder en los Estados Unidos, entonces el algoritmo está intencionalmente sesgado [61].

Dentro de la inclusión existen tres aspectos específicos: equidad, no discriminación y neutralidad, que son conceptos relacionados, pero que tienen matices diferentes:

4.1 Equidad

Se refiere a la justicia y la imparcialidad en el tratamiento de las personas y en la toma de decisiones. Esto incluye asegurar que todos tengan acceso a oportunidades y recursos de manera justa y sin discriminación. Los sistemas de IA deben tratar a los individuos sin distinción, independientemente de sus características [42].

Equidad también significa que todas las personas tienen los mismos derechos y deben tener acceso a las mismas oportunidades. La igualdad es el principio que reconoce que todas las personas son semejantes ante la ley.

4.2 No discriminación

La no discriminación es el principio que prohíbe tratar de forma desfavorable o que atenten en contra de la dignidad humana de una persona por motivos de su raza, orientación sexual, identidad de género, edad, discapacidad, religión, nacionalidad, ideología o afinidad política [42]. Implica tratar igual a todas las personas sin hacer distinciones injustas o discriminatorias, reconociendo el valor intrínseco de cada individuo.

4.3 Neutralidad

Se refiere a la imparcialidad o ausencia de prejuicios o sesgos en la toma de decisiones. Esto se traduce en asegurar que las decisiones se basen en hechos y no en ideas preconcebidas o creencias personales [43].

Los tres conceptos: equidad, no discriminación y neutralidad están relacionados e influyen el uno en el otro. Por ejemplo, la equidad puede requerir la no discriminación y la neutralidad, y la no discriminación puede requerir la equidad y la neutralidad.

La equidad es el principio que busca compensar las desigualdades históricas o estructurales que afectan a ciertos grupos sociales, otorgándoles un trato preferente o diferenciado para garantizar su igualdad real de oportunidades y resultados. Hay que asegurar que el sistema de IA trate a todas las personas de manera justa y equitativa, y que no discrimine por razones como la raza, sexo, identidad de género, edad, discapacidad, creencias religiosas o pertenencia a algún grupo minoritario [44] [45].

La neutralidad, por su parte, es el principio que implica abstenerse de tomar partido o favorecer a una parte en un conflicto o situación, actuando con imparcialidad y objetividad.

Para evitar sesgos, se puede seguir las siguientes recomendaciones:

- Contar con equipos diversos que representen a diferentes grupos sociales y puntos de vista.

- Usar datos no contaminados para entrenar a la IA, eliminando o equilibrando las variables que puedan causar discriminación.
- Revisar periódicamente los algoritmos para detectar y corregir posibles errores o desviaciones.
- Mayor entendimiento de lo que implica la inteligencia artificial y sus limitaciones, así como de los objetivos y valores que se quieren alcanzar con ella.
- Formar a las personas que diseñan, implementan o supervisan los algoritmos para que sepan que es un sesgo inconsciente [46].

La neutralidad significa actuar con objetividad y justicia al decidir algo. Esto implica que las decisiones se fundamenten en los hechos y no en opiniones o juicios preconcebidos.

5. Privacidad

Es importante proteger la privacidad y garantizar que la IA no abuse del acceso a los datos personales. El respeto a la privacidad se refiere a la necesidad de proteger la información de los usuarios y de asegurar que no se haga mal uso de ella [34]. Esto incluye considerar aspectos como:

La recopilación de datos personales de los usuarios debe hacerse solo cuando sea estrictamente necesario, previo consentimiento; que se utilicen para los fines para los que se han obtenido; que no se vendan a terceros; que estén protegidos contra el acceso no autorizado y contra el robo o la pérdida de datos; y proporcionar a los usuarios información clara y comprensible sobre cómo se recopilan, utilizan y resguardan dichos datos personales [47].

Los datos personales, como ya se ha dicho en otro apartado, son, entre otros: domicilio particular, RFC o número de contribuyente, ingresos, números de cuenta, hijos, cónyuge, familia, escuela, vehículo, placas, padecimientos físicos o mentales, historial clínico, adicciones, asuntos legales, calificaciones escolares, problemas familiares, ubicación geográfica, agenda, edad, sexo, nombre de usuario, búsquedas en Internet e intereses, por mencionar solo algunos. Las contraseñas y el número de identificación personal (NIP) se consideran datos confidenciales [48].

6. Sostenibilidad

Los sistemas de IA deben diseñarse para ser sostenibles y garantizar que no perjudiquen al medio ambiente ni a la sociedad. Es importante que la IA tenga en cuenta las consecuencias sociales y medioambientales de sus decisiones y acciones

a largo plazo [49]. La sostenibilidad se refiere a la capacidad de una actividad o sistema para permanecer en el tiempo y para satisfacer las necesidades del presente sin comprometer la capacidad de las generaciones futuras para cubrir sus propias necesidades. En el contexto del software, la sostenibilidad es la eficiencia energética para asegurar que sea eficaz en el uso de energía y recursos, para minimizar el impacto ambiental y maximizar la durabilidad.

La escalabilidad significa que el sistema pueda adaptarse y actualizarse de manera eficiente para satisfacer las necesidades de los usuarios y los cambios en el entorno [50].

Es importante considerar las implicaciones a largo plazo al diseñar y utilizar la tecnología de la IA porque también tiene el potencial de utilizar muchos recursos y dañar el medio ambiente debido a su elevado consumo de energía y a su huella de carbono. Este principio subraya la necesidad de una investigación ética de la IA en cuanto a su operación y rendimiento por lo que es necesario establecer reglamentos y normas claras y aplicables para el desarrollo, el uso y la eliminación de las tecnologías de IA.

Menos, es más

Hasta aquí se han enumerado principios que deben observarse para asegurar que la IA sea responsable, como la equidad, la no discriminación, la neutralidad, la transparencia, la seguridad, la protección de la privacidad y la sostenibilidad.

Es crucial que los desarrolladores de sistemas de IA asuman la responsabilidad de las acciones y decisiones tomadas por los softwares que crean. Deben dar prioridad a la transparencia en la recogida, uso y protección de los datos personales de los individuos, así como en el funcionamiento de sus sistemas de IA. Los desarrolladores deben responsabilizarse de las decisiones y ser capaces de ofrecer explicaciones exhaustivas al respecto. Garantizar la transparencia en la utilización de los datos y el funcionamiento es esencial para mantener la confianza de los usuarios.

El reto es cómo fomentar que los desarrolladores de IA visualicen las consecuencias de sus acciones, de sus decisiones, y actúen de acuerdo con principios universales éticos [51]. Cuando se habla de ética y responsabilidad en el desarrollo y uso de IA existen una serie de principios y aspectos clave:

1. Honradez

La honradez es un principio fundamental de ética en cualquier contexto, incluyendo el desarrollo y uso de IA. Implica ser honesto y transparente en acciones y decisiones toda vez que es fundamental establecer confianza y credibilidad de los sistemas de IA [52].

En este contexto, la honradez puede incluir: decir las limitaciones y los posibles riesgos del sistema de IA; ser transparente sobre cómo se recopilan, utilizan y protegen los datos personales; decir cómo es el funcionamiento del sistema de IA y las decisiones que toma, y evitar la manipulación intencional o el uso indebido del sistema para perjudicar a otros.

Puede verse de esta manera que la honradez es un principio fundamental para garantizar que la IA se utilice de manera adecuada y establecer la confianza de los usuarios en la tecnología.

2. Intencionalidad

Es posible extender el principio de la honradez al problema de la intencionalidad. La intencionalidad se refiere a la capacidad de un agente, ya sea una persona o un sistema de IA, para planear su estrategia y actuar de acuerdo con ella [53]. En el contexto de la IA, la intencionalidad solo reside en el desarrollador e incluye los siguientes aspectos:

- Responsabilidad por las acciones y decisiones para poder ser capaz de responder por ellas.
- Ser transparente sobre las intenciones y los objetivos del sistema de IA y sobre cómo toma sus decisiones.
- Ser honesto sobre las intenciones y los objetivos del sistema de IA.
- Ser íntegro es actuar de manera honesta y justa y que no sea manipulado o utilizado el sistema de forma indebida [54].

La intencionalidad desempeña un papel crucial en el desarrollo y el uso de los sistemas de IA. Hace énfasis en la necesidad de garantizar que los desarrolladores sean responsables, transparentes, honestos y actúen con integridad en sus decisiones y acciones.

3. Tribunal de la conciencia

El concepto kantiano del tribunal de la conciencia se refiere a la idea de que cada persona tiene una conciencia interna que actúa como juez, que evalúa sus acciones y decisiones y que le permite distinguir entre lo correcto y lo incorrecto. Según Kant, esta conciencia es el fundamento de la moralidad y obliga a la persona a actuar de acuerdo con principios universales que son aplicables a todos los seres racionales [55].

En el contexto de la IA, el concepto del tribunal de la conciencia puede aplicarse para las siguientes funciones:

- Evaluar el comportamiento y las decisiones de los sistemas. Si cumplen con los principios universales morales del bien común, y si son aceptables para cualquier persona [56].
- Fomentar la responsabilidad y la reflexión en el desarrollo y uso de la IA. Hay que considerar que los desarrolladores actúen de acuerdo con los principios de no maleficencia y beneficencia y reflexionen sobre las posibles consecuencias de sus algoritmos.
- Promover la honestidad en el desarrollo y uso de la IA. Debe existir transparencia sobre las intenciones y objetivos de cada proyecto. Las directrices deben considerar tanto los beneficios como los riesgos potenciales de la tecnología, esto significa que los desarrolladores y usuarios deben ser conscientes de cómo sus acciones podrían tener un impacto dual en la sociedad, y garantizar que la IA se desarrolle y utilice de manera responsable.

Un imperativo categórico para Kant es examinar si la acción puede, por ejemplo, aplicarse de modo tal que sirviera de ejemplo o ley para el resto de las personas. Si la acción no puede ser generalizada, entonces no debe realizarse [62].

Conclusión

Menos es más significa reconstruir una ética clásica normativa para un futuro responsable de la IA desechando principios abstractos y concentrarse en acciones concretas de mejora continua. Para prevenir los usos nocivos de la IA se propone que sean necesarias la regulación y la supervisión. Los gobiernos y los organismos reguladores pueden establecer leyes y reglamentos que rijan el desarrollo y el uso de la IA, y garantizar que la tecnología se utilice de forma coherente con los principios éticos aquí planteados. Esto implicaría supervisar el uso de la IA y también

aplicar sanciones a quienes violen estos principios. La creación objetiva y positiva de un marco regulador es inaplazable.

El desarrollo y el uso de la IA deben guiarse por principios éticos, centrándose en equilibrar sus beneficios y riesgos potenciales. La regulación, aunque a muchos les moleste, y la supervisión, sin cortar la creatividad y otorgando los espacios de secreto industrial, son necesarias para garantizar que la IA se desarrolle y utilice de forma ética, y evitar así usos nocivos que puedan tener repercusiones perniciosas de largo plazo en la sociedad.

Fomentar la colaboración y el diálogo entre investigadores, desarrolladores y usuarios de la IA de forma interdisciplinaria, compartiendo conocimientos y buenas prácticas, puede ayudar a garantizar que la tecnología se utilice de forma que beneficie a la humanidad.

Las normas y orientaciones para el desarrollo y uso de la IA deben tener como objetivo maximizar los beneficios de la IA y minimizar los daños potenciales. Una parte fundamental de esto es la supervisión y la gobernanza. Las instituciones gubernamentales y otras organizaciones deben establecer directrices y políticas responsables para la IA. Supervisar que los sistemas se diseñen y utilicen de forma ética y respetuosa con los derechos, la dignidad y el bienestar de las personas. Con leyes, regulaciones, restricciones y supervisión adecuadas, se puede ayudar a garantizar que la tecnología esté al servicio de la humanidad. La IA tiene un gran potencial para mejorar la vida de las personas, pero solo si se desarrolla de la forma correcta.

Se debe dar prioridad a valores como la equidad, la transparencia y la responsabilidad para cosechar todos los beneficios de la IA de una manera segura y ética. Es menester fomentar la colaboración y el diálogo entre investigadores, desarrolladores y partes interesadas, trabajando juntos y compartiendo conocimientos y buenas prácticas se puede ayudar a garantizar que la tecnología de la IA se utilice de forma que beneficie a la humanidad, minimizando al mismo tiempo la posibilidad de usos perjudiciales.

Las visiones apocalípticas de un futuro controlado por inteligencias artificiales malvadas distorsionan el verdadero potencial de ayuda que los modelos de aprendizaje automático puedan tener y de los que, incluso, se puede aprender gracias a la sistematización perfecta, tanto de la información como de las tareas [57]. Los mayores riesgos no estriban en que robots del futuro tomen un control dictatorial y físico sobre las personas, sino en que sistemas superinteligentes (ASI - Artificial

Super Intelligence) comiencen a incidir en la toma de decisiones de las personas, sin que los afectados siquiera sean conscientes de esta alienación.

La IA puede ser una copilota, aliada, compañera, asesora y tutora en este largo, pero vertiginoso viaje rumbo al desarrollo tecnológico-humano. Por tanto, no puede ser vista la creación como algo independiente de su creador y si el ser humano es capaz de construir una tecnología que *lo rebase*, también se debe considerar que él es el único responsable. Por ello, por más autonomía que la IA vaya escalando, la ética no es una atribución para un ente no-humano, sino del propio hombre como sujeto de responsabilidad. El futuro de los seres humanos como especie dependerá de las decisiones morales que tomen los desarrolladores individuales y de las decisiones colectivas que tome la sociedad en relación con la IA. Estas decisiones determinarán si la IA le otorga poder o la destruye.

Es importante que los desarrolladores tomen internamente decisiones éticas al diseñar sistemas de IA y que la sociedad establezca normativas que rijan el desarrollo y el uso de la IA de forma responsable. Si se tiene en cuenta las implicaciones morales y se aplica la normativa adecuada, se puede garantizar que la IA beneficie a la humanidad en lugar de suponer una amenaza.

En conclusión, reconstruir una ética reguladora clásica para un futuro responsable de la IA basada en la honestidad, las buenas intenciones y la conciencia moral representa un paso adelante crucial para garantizar que la IA se desarrolle y despliegue de forma coherente con los valores humanos y el bien común. No es una aspiración cándidamente optimista. En última instancia, depende de todos, como individuos y miembros de la sociedad dar prioridad a la honestidad, las buenas intenciones y análisis férreo de la brújula moral a la hora de guiar el desarrollo y el despliegue de la IA. De este modo, se puede crear un futuro en el que se respete los valores humanos y éticos comunes.

Difícilmente habrá pasos atrás en el desarrollo y uso de las herramientas de la inteligencia artificial en todos los campos de acción del ser humano, pero también depende del hombre, como especie, que sus inventos no sean perjudiciales para el conjunto de la sociedad. Al ser reflexivo, compasivo y un administrador vigilante del progreso, el hombre y la IA pueden desarrollar y aplicar de manera profundamente enriquecedora y edificante nuevas soluciones. Pero los tomadores de decisiones de gran alcance deben tener buenas intenciones y estar decididos a dar forma al futuro de esta tecnología. Con veracidad, responsabilidad y atención a los valores humanos compartidos, se puede crear un mundo basado en IA que sea mayor que la suma de sus partes, humana y mecánica. El futuro aún no está escrito, y depende

del ser humano guiar estas poderosas herramientas para ayudar a crear una historia mejor y más brillante para todos. Se tiene la oportunidad, ahora hay que encontrar la voluntad y la decisión.

Referencias

[1] Asimov, "I, Robot." New York, NY: Gnome Press, 1950.

[2] P. Ghosh, "La premisa única que reescribe las leyes de la robótica de Isaac Asimov, el padre de la ciencia ficción," BBC.com. Acceso jun. 2023. [En línea] Disponible: https://www.bbc.com/mundo/noticias-40446863.

[3] J. McCarthy, et al., "A proposal for the Dartmouth Summer Research Project on Artificial Intelligence," 1955. Stanford.edu. Acceso jun. 2023. [En línea] Disponible: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

[4] C. Bartneck, C. Lütge, A. Wagner, y S. Welsh, "Responsibility and Liability in the Case of AI Systems," en *An Introduction to Ethics in Robotics and AI*, C. Bartneck, et al., Eds. Springer International Publishing, 2021, pp. 39-44.

[5] C. Bartneck, C. Lütge, A. Wagner, y S. Welsh, "Trust and Fairness in AI Systems," en *An Introduction to Ethics in Robotics and AI*, Springer Briefs in Ethics, Cham: Springer, 2021. [Online]. Available: https://doi.org/10.1007/978-3-030-51110-4_4.

[6] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo y L. Floridi, "The ethics of algorithms: key problems and solutions," *AI & Soc.*, vol. 37, no. 1, pp. 215-230, Feb. 2022. doi: 10.1007/s00146-021-01154-8.

[7] H. Roberts, J. Cowls, E. Hine, A. Tsamados, M. Taddeo y L. Floridi., "Achieving a 'Good AI Society': Comparing the Aims and Progress of the EU and the US," *Sci. Eng. Ethics*, vol. 27, no. 2, pp. 68, 2021, doi: 10.1007/s11948-021-00340-7.

[8] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, y L. Floridi, "Artificial Intelligence and the 'Good Society': the US, EU, and UK approach," *Sci. and Eng. Ethics*, vol. 24, no. 2, pp. 505-528, Feb. 2018, doi: 10.1007/s11948-017-9901-7.

[9] M. Dastani y V. Yazdanpanah, "Responsibility of AI Systems," *AI & Soc.*, vol. 38, pp. 843–852, jun. 2022. doi: 10.1007/s00146-022-01481-4.

[10] M. Taddeo y A. Blanchard, "Accepting Moral Responsibility for the Actions of Autonomous Weapons Systems—a Moral Gambit," *Philos. Technol.,* vol. 35, no. 1, pp. 78, 2022, doi: 10.1007/s13347-022-00571-x

[11] Hayes, "Speculative Stock: Definition, Uses, Sector Examples," Investopedia.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/b30

[12] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Sci. Eng. Ethics*, vol. 26, no. 6, pp. 2141-2168, 2020, doi: 10.1007/s11948-019-00165-5.

[13] F. Morandín-Ahuerma, "Neuroética fundamental y teoría de las decisiones." 2021, Puebla, México: Consejo de Ciencia y Tecnología del Estado de Puebla (Concytep).

[14] F. Morandín-Ahuerma, "Causalidad bivalente en la toma de decisiones morales," en *Neuroética fundamental y teoría de las decisiones*, 2021, Consejo de Ciencia y Tecnología del Estado de Puebla (Concytep), pp. 33-42.

[15] IEEE, "7000-2021 - IEEE Standard Model Process for Addressing Ethical Concerns during System Design," Investopedia.com. Acceso jun. 2023. [En línea] Disponible: https://ieeexplore.ieee.org/document/9536679

[16] M. Taddeo y L. Floridi, "How AI Can Be a Force for Good – An Ethical Framework to Harness the Potential of AI While Keeping Humans in Control," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed. Cham: Springer International Publishing, 2021, pp. 91-96.

[17] J. Fjeld, N. Achten, H. Hilligoss, A. C. Nagy, y M. Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI," Berkman Klein Center for Internet & Society, Research Publication No. 2020-1, 2020, doi: 10.2139/ssrn.3518482.

[18] A.J. Andreotta, N. Kirkham, y M. Rizzi, "AI, big data, and the future of consent," *AI & Soc.*, vol. 37, no. 4, pp. 1715-1728, 2022, doi: 10.1007/s00146-021-01262-5

[19] J. Mökande, J. Morley, M. Taddeo y L. Floridi, "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations," *Sci. Eng. Ethics*, vol. 27, no. 4, p. 44, 2021, doi: 10.1007/s11948-021-00319-4.

[20] N. Diakopoulos, "Transparency. Accountability, Transparency, and Algorithms," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford University Press, 2020, pp. 197-213.

[21] J. Kroll, "Accountability in Computer Systems," en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford University Press, 2020, pp. 180–196.

[22] M. Taddeo, T. McCutcheon, y L. Floridi, "Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword," en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed. Cham: Springer International Publishing, 2021, pp. 289-297.

[23] L. Ibarra, D. Balderas, P. Ponce & A. Molina, "Fast Execution of Black-Box Algorithms Through a Piece-Wise Linear Interpolation Technique," *Arab. J. Sci. Eng.*, vol. 44, no. 11, pp. 9443-9453, 2019. Disponible en: https://doi.org/10.1007/s13369-019-04042-y

[24] B. Doerr, C. Doerr, y F. Ebel, "From black-box complexity to designing new genetic algorithms," *Theor. Comput. Sci.*, vol. 567, pp. 87-104, 2015, doi: 10.1016/j.tcs.2014.11.028

[25] S. Russell y P. Norvig, "Philosophy, ethics, and safety of AI," en *Artificial Intelligence: A Modern Approach*, Londres: Pearson, 2022, pp. 1032-1062.

[26] D.M. Monte-Serrat y C. Cattani, "The natural language for artificial intelligence." Cambridge, MA, USA: Academic Press, 2021.

[27] J.L. Gastaldi, "Why can computers understand natural language? the structuralist image of language behind word embeddings," *Phil. & Tech.*, vol. 34, no. 1, pp. 149-214, 2021.

[28] M.J. McGowan, “The rise of computerized high-frequency trading: use and controversy,” *Duke L. & Tech. Rev.*, vol. 9, p. 1, 2010.

[29] Cioroianu, S. Corbet, y C. Larkin, “Guilt through association: Reputational contagion and the Boeing 737-MAX disasters,” *Economics Letters*, vol. 198, p. 109657, 2021.

[30] CNN en Español, “Ethiopian Airlines: El piloto del vuelo 302 tuvo problemas de control de vuelo,” CNN.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/ci9

[31] R. Engineering, “NASA’s 150 Million Dollar Coding Error,” [Video] 2018; Disponible en: https://youtu.be/CkOOazEJcUc

[32] J.M. Durán y K.R. Jongsma, “Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI,” *Journal of Medical Ethics*, vol. 47, no. 5, pp. 329-335, 2021. https://jme.bmj.com/content/47/5/329

[33] L. Cotino Hueso y J. Castellanos Claramunt, “Transparencia y explicabilidad de la inteligencia artificial y ‘compañía’ Para qué, para quién y cuánta,” Tirant lo Blanch, 2022.

[34] C. Véliz, “Privacy Is Power: Why and How You Should Take Back Control of Your.” NY: Penguin Random House, 2022.

[35] K.A. Chagal-Feferkorn, “Am I an algorithm or a product: when products liability should apply to algorithmic decision-makers,” *Stan. L. & Poly. Rev.*, vol. 30, p. 61, 2019.

[36] K. Amer y J. Noujaim, “The Great Hack [Nada es privado],” Netflix, Estados Unidos, 2019.

[37] M. Wazid, S. Zeadally, y A.K. Das, “Mobile banking: evolution and threats: malware threats and security solutions,” *IEEE Consum. Electron.*, vol. 8, no. 2, pp. 56-60, 2019, doi: 10.1109/MCE.2018.2881291.

[38] S. Sambangi and L. Gondi, “A machine learning approach for DDoS (distributed denial of service) attack detection using multiple linear regression.” Suiza: MDPI, 2020.

[39] SBA, “Strengthen your cybersecurity,” SBA.gov. Acceso jun. 2023. [En línea] Disponible: https://www.sba.gov/business-guide/manage-your-business/strengthen-your-cybersecurity.

[40] F.J. Zuiderveen Borgesius, “Strengthening legal protection against discrimination by algorithms and artificial intelligence,” *J. Hum. Rights*, vol. 24, no. 10, pp. 1572-1593, 2020.

[41] T. Gebru, “Race and Gender,” in *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds. Oxford University Press, 2020, pp. 252–269.

[42] J. W. Gichoya, L. G. McCoy, L. A. Celi, y M. Ghassemi, “Equity in essence: a call for operationalising fairness in machine learning for healthcare,” *BMJ Health Care Inform.*, vol. 28, no. 1, pp. e100289, 2021, doi: 10.1136/bmjhci-2020-100289.

[43] M. Phillips-Brown, “Algorithmic neutrality,” 2023, https://arxiv.org/abs/2303.05103

[44] V. Durrer, T. Miller, L. A. Celi, and M. Ghassemi, “The Routledge Handbook of Global Cultural Policy,” 1st ed. Abingdon: Routledge, 2018.

[45] M.L. Stefano y P. Davis, “The Routledge guide to intangible cultural heritage.” Routledge, 2017.

[46] Y. Kim, J. Kim, S. Kim, y S. Lee, “Lipschitz continuous autoencoders in application to anomaly detection.” Proceedings of Machine Learning Research, 2020.

[47] J.P. Choi, D.-S. Jeon, y B.-C. Kim, “Privacy and personal data collection with information externalities,” *Journal of Public Economics*, vol. 173, pp. 113-124, 2019.

[48] V. Jesus y S. Mustare, “I did not accept that: Demonstrating consent in online collection of personal data,” Springer, 2019.

[49] B.C. Stahl, “Artificial Intelligence for a Better Future. An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies,” Springer International Publishing, 2021.

[50] N. Beigi-Mohammadi, et al., “On efficiency and scalability of software-defined infrastructure for adaptive applications,” IEEE, 2016.

[51] T. Hagendorff, “The Ethics of AI Ethics: An Evaluation of Guidelines,” Minds and Machines, vol. 30, no. 1, pp. 99-120, 2020.

[52] W. Fleeson, et al., “Honesty as a trait,” Current Opinion in Psychology, vol. 2022, p. 101418.

[53] J. Kinkaid, “Phenomenology, idealism, and the legacy of Kant,” *Br. J. Hist. Philos.*, vol. 27, no. 3, pp. 593-614, 2019.

[54] J.T. Hancock, M. Naaman, y K. Levy, “AI-mediated communication: Definition, research agenda, and ethical considerations,” *J. Comput. Mediat. Commun.*, vol. 25, no. 1, pp. 89-100, 2020.

[55] J.D.M. Derrett, “Justice, equity and good conscience,” in Changing law in developing countries, Routledge, 2021, pp. 114-153.

[56] F. Morandín-Ahuerma, A. Romero-Fernández, L. Villanueva-Méndez, y E. Santos-Cabañas, “Hacia una fundamentación ético-normativa del sujeto de derecho,” *Rev. Jurid. Crítica y Der.*, vol. 4, no. 2, pp. 1-12, Ene. 2023, doi: 10.29166/cyd.v4i6.4242.

[57] C. Bartneck, C. Lütge, A. Wagner y S. Welsh, “Psychological Aspects of AI,” in *An Introduction to Ethics in Robotics and AI*, Springer International Publishing, 2021, pp. 55-60.

[58] C. Molnar, “Interpretable Machine Learning: A Guide for Making Black Box Models Explainable,” Independently published, 2019.

[59] E. Feldman, “Are A.I. Image Generators Violating Copyright Laws?” *Smithsonian Magazine*. Smithsonianmag.com. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/cja

[60] B. Hunting, “A Timeline of the NHTSA Investigation Into Tesla Autopilot and Full Self-Driving Technology,” CapitalOne. Acceso jun. 2023. [En línea] Disponible: https://bsu.buap.mx/cm0

[61] Citizen.org, “Report: Algorithms Are Worsening Racism, Bias, Discrimination,” Citizen.org Acceso jun. 2023. [En línea] Disponible: https://www.citizen.org/news/report-algorithms-are-worsening-racism-bias-discrimination/

[62] E. Kant, “Grundlegung zur Metaphysik der Sitten.” (Fundamentos de la metafísica de la moral). Jazzybee Verlag, 2012.

Glosario

Algoritmo: Secuencia finita de instrucciones bien definidas que puede ejecutar un ordenador para resolver un problema concreto o realizar una tarea específica. Un algoritmo de IA es un conjunto de reglas y procedimientos diseñados para permitir a un programa informático aprender de los datos y hacer predicciones, decisiones o recomendaciones basadas en ese aprendizaje. Estos algoritmos suelen implicar complejos modelos matemáticos y técnicas estadísticas. Los algoritmos de IA pueden utilizarse para una amplia gama de aplicaciones, como el reconocimiento de imágenes, el procesamiento del lenguaje natural, los vehículos autónomos, entre otras aplicaciones.

Algoritmo de caja negra: Modelo de aprendizaje automático u otro sistema de IA cuyo funcionamiento interno no es fácilmente comprensible o interpretable por los humanos. Concretamente es difícil o imposible discernir cómo el algoritmo toma decisiones o llega a su resultado. Este suele ser el caso de los modelos complejos de aprendizaje automático que tienen muchas capas, nodos y conexiones, lo que dificulta trazar el camino de la información a través del sistema. Los algoritmos de caja negra se utilizan mucho en aplicaciones de IA como el reconocimiento de imágenes, el procesamiento del lenguaje natural y los modelos predictivos, entre otros. Sin embargo, su falta de transparencia ha suscitado preocupación por su potencial de sesgo y discriminación, así como por su responsabilidad en contextos críticos de toma de decisiones como la sanidad, las finanzas y la justicia penal.

Aprendizaje automático (machine learning): Es una rama de la inteligencia artificial que se centra en la creación de algoritmos informáticos que puedan aprender por sí mismos de los datos recibidos y hacer predicciones o tomar decisiones basadas en ellos. En lugar de estar explícitamente programados para realizar una tarea específica, estos algoritmos utilizan métodos estadísticos y modelos matemáticos para aprender patrones y relaciones a partir de grandes conjuntos de datos. El objetivo del ML es desarrollar algoritmos capaces de mejorar su rendimiento con el tiempo a medida que reciben más y más datos con cierta autonomía, lo que les permite adaptarse a entornos cambiantes y hacer predicciones o tomar decisiones más precisas. El ML tiene numerosas aplicaciones en campos como la visión por computadora, el procesamiento del lenguaje natural, la robótica y las finanzas, entre otros.

Aprendizaje profundo (deep learning): Es un subconjunto del aprendizaje automático que consiste en entrenar redes neuronales artificiales con múltiples capas para aprender y hacer predicciones o tomar decisiones basadas en datos previos.

El aprendizaje profundo es un tipo de aprendizaje automático que utiliza redes neuronales artificiales con muchas capas (neuronas artificiales) para aprender automáticamente patrones y relaciones en los datos. Esto permite crear modelos sofisticados capaces de reconocer y clasificar patrones complejos en imágenes, sonidos y textos. Se ha utilizado en muchas aplicaciones, como los coches de conducción autónoma y el reconocimiento de voz.

Aprendizaje por refuerzo: Es un tipo de aprendizaje automático en el que un agente aprende a tomar decisiones recibiendo retroalimentación en forma de recompensas o castigos. El agente prueba diferentes acciones para aprender cuáles le reportan la mayor recompensa y, con el tiempo, aprende a realizar las mejores acciones para alcanzar sus objetivos. El aprendizaje por refuerzo se ha utilizado para entrenar robots en la realización de tareas y para enseñar a los agentes a jugar a juegos como Go y ajedrez a niveles sobrehumanos. El principal reto es equilibrar la exploración de nuevas acciones con la explotación de lo que el agente ya sabe para maximizar su recompensa.

Aprendizaje supervisado: Consiste en entrenar un modelo con resultados o etiquetas correctas. A continuación, el modelo aprende a predecir el resultado ideal para datos nuevos que no haya visto. El aprendizaje no supervisado, por el contrario, consiste en dejar que encuentre patrones y relaciones en datos no etiquetados. Un ejemplo habitual de aprendizaje supervisado es la clasificación de imágenes, cuyo objetivo es clasificar una imagen en una de varias categorías predefinidas como “gato” o “perro”. En cambio, en el no supervisado, el algoritmo aprende a encontrar patrones y relaciones en los datos sin etiquetas ni categorías predefinidas, por ejemplo, para hacer sugerencias en las plataformas como Netflix.

Armas autónomas: También conocidas como sistemas de armas autónomas letales (LAWS, por sus siglas en inglés), se refieren a armas que pueden seleccionar y atacar objetivos sin intervención humana. Estas armas pueden funcionar por sí solas, basándose en instrucciones preprogramadas o mediante algoritmos de inteligencia artificial (IA).

Auto-mejora recursiva: Se refiere a un proceso en el que las máquinas se mejoran a sí mismas al repetir y aplicar indefinidamente un algoritmo de ensayo y error autogestivo, utilizando el conocimiento y los medios de mejora generados durante iteraciones anteriores del proceso. Por ejemplo, los programas de inteligencia artificial podrían mejorar sus propios algoritmos, los sensores podrían desarrollar cada vez más capacidades para percibir e interpretar su entorno y los sistemas robóticos podrían construir robots cada vez más sofisticados basándose en habilidades

y componentes desarrollados en generaciones anteriores. A medida que las máquinas se vuelvan más inteligentes y hábiles a través de esta auto-mejora recursiva, podría conducir a un aumento exponencial de sus capacidades y a mejoras inimaginables en las tecnologías que desarrollan, lo que potencialmente daría lugar a lo que se conoce como una singularidad tecnológica lo que para muchos es considerado un verdadero peligro.

Big data: Hace referencia a conjuntos de datos extremadamente grandes y complejos que no pueden procesarse con las técnicas tradicionales de tratamiento de datos. Estos conjuntos de datos suelen caracterizarse por su volumen, velocidad y variedad. El término "big data" se ha hecho cada vez más popular en los últimos años debido al crecimiento exponencial de los datos generados por diversas fuentes, como las redes sociales, las transacciones en línea, la investigación científica y otras fuentes. Estos datos suelen estar desestructurados o semiestructurados, lo que dificulta su análisis y la extracción de información con los métodos tradicionales.

Capas: Se refiere a un conjunto de neuronas artificiales que procesan un tipo específico de datos de entrada y producen un conjunto correspondiente de salidas. Las redes neuronales artificiales, que son un tipo de modelo de aprendizaje automático inspirado en la estructura del cerebro humano, suelen estar compuestas por varias capas de neuronas artificiales. Cada capa recibe datos de entrada de la capa anterior, los procesa utilizando un conjunto de pesos y sesgos, y produce una salida que pasa a la capa siguiente. Hay diversos tipos de capas que se utilizan habitualmente en las redes neuronales: capas de entrada, capas ocultas y capas de salida. Las capas de entrada reciben los datos brutos de entrada, mientras que las capas ocultas realizan el procesamiento intermedio y la extracción de características. Las capas de salida producen el resultado final del modelo, como una predicción de clasificación o regresión. La estructura y la configuración de las capas de una red neuronal pueden tener un impacto significativo en el rendimiento del modelo, incluida su precisión, velocidad y capacidad de generalización a nuevos datos. Los investigadores y profesionales experimentan a menudo con distintas arquitecturas y configuraciones de capas para optimizar el rendimiento de sus modelos en una tarea determinada.

GPT (Transformador preentrenado generativo): Se refiere a una familia de modelos de lenguaje de Inteligencia Artificial desarrollados por la empresa OpenAI que utilizan una arquitectura de transformador para aprender a producir texto coherente y relevante en una variedad de tareas de procesamiento del lenguaje natural. Estos modelos son "pre-entrenados" en grandes cantidades de texto sin

etiquetar, lo que les permite aprender patrones y estructuras lingüísticas complejas. Luego, se pueden ajustar finamente para tareas específicas de procesamiento del lenguaje natural, como la traducción automática, la generación de resúmenes y la respuesta automática a preguntas concretas. También está siendo utilizado en una amplia variedad de aplicaciones, desde chatbots de atención al cliente hasta la generación de poesía, narrativa, programación y música, entre otras.

Minería de datos: En la jerga de la IA, la minería de datos se refiere al proceso de extraer patrones y conocimientos de grandes conjuntos de datos mediante técnicas informáticas avanzadas como el aprendizaje automático, el análisis estadístico y el reconocimiento de patrones. La minería de datos se utiliza a menudo para descubrir patrones ocultos, relaciones y correlaciones en conjuntos de datos complejos que pueden ser difíciles o imposibles de identificar utilizando métodos tradicionales de análisis de datos. Consiste en aplicar diversos algoritmos y técnicas a grandes conjuntos de datos para identificar tendencias y patrones que puedan servir de base para la toma de decisiones empresariales, la investigación científica y otras aplicaciones. La minería de datos puede aplicarse a una amplia gama de conjuntos de datos, incluidos los estructurados, semiestructurados y no estructurados. Algunas aplicaciones habituales de la minería de datos son el análisis de cestas de mercado, la detección de fraudes, la segmentación de clientes y el análisis predictivo. Aunque la minería de datos puede ser una poderosa herramienta para generar ideas y fundamentar la toma de decisiones, también plantea problemas en torno a la privacidad, la seguridad y el uso ético de los datos. Por ello, la minería de datos suele estar sujeta a regulación y supervisión, sobre todo en sectores como la sanidad y las finanzas, donde el uso de datos personales está muy regulado.

Inteligencia artificial: Un sistema basado en una máquina que puede hacer predicciones, recomendaciones o tomar decisiones que influyan en entornos reales o virtuales, para un conjunto dado de objetivos definidos por el ser humano. Los sistemas de inteligencia artificial utilizan entradas basadas en máquinas y humanos para: percibir entornos reales y virtuales; abstraer dichas percepciones en modelos mediante análisis de forma automatizada y utilizar la inferencia de modelos para formular opciones de información o acción.

Sesgos: Se refiere a la presencia de errores o imprecisiones sistemáticos en un modelo de aprendizaje automático que pueden dar lugar a resultados injustos o discriminatorios. El sesgo puede surgir de varias maneras, como por ejemplo del diseño del modelo, de la calidad o representatividad de los datos de entrenamiento, o de las suposiciones y preferencias de las personas implicadas en el desarrollo o

uso del modelo. Por ejemplo, un sistema de reconocimiento facial que ha sido entrenado en un conjunto de datos que contiene principalmente rostros de hombres blancos puede ser menos preciso a la hora de reconocer rostros de mujeres o de personas de otros orígenes raciales o étnicos. Este es un ejemplo de sesgo algorítmico, en el que el modelo ha aprendido a asociar ciertas características con determinados grupos de personas, y es incapaz de generalizar a otros grupos. El sesgo en la IA puede tener graves consecuencias, sobre todo en aplicaciones que tienen un impacto significativo en la vida de las personas, como la contratación, los préstamos y la justicia penal. La IA sesgada puede dar lugar a un trato injusto, perpetuar las desigualdades sociales existentes y socavar la confianza pública en la tecnología. Para abordar el sesgo en la IA, investigadores y profesionales están desarrollando una serie de métodos y técnicas, como métricas de imparcialidad, auditorías algorítmicas y estrategias de mitigación del sesgo. Estos enfoques pretenden identificar y mitigar el sesgo en los modelos de aprendizaje automático y promover una IA más equitativa y ética.

Web profunda (Deep Web): La web profunda es la parte de Internet no indexada por los motores de búsqueda y a la que no se puede acceder fácilmente a través de los navegadores habituales. Consiste en sitios web y contenidos que no están destinados a ser de acceso público. Para entender la web profunda, es útil pensar en Internet como en un iceberg. La web superficial, con la que la mayoría de la gente interactúa a diario, representa la parte visible del iceberg. Incluye sitios web, resultados de motores de búsqueda, plataformas de redes sociales y otros contenidos de acceso público que pueden ser indexados por Google, Bing, Opera, etcétera. Por otro lado, la web profunda representa la parte sumergida del iceberg. Esto incluye bases de datos privadas, sitios web protegidos por contraseña, documentos gubernamentales, información confidencial y otros contenidos que requieren credenciales específicas, permisos o software especializado para acceder a ellos. Contrariamente a la creencia popular, la *deep web* no es intrínsecamente ilegal o malévola. Existe principalmente para proteger información sensible y proporcionar acceso restringido a personas u organizaciones autorizadas, por ejemplo, en países en que no existe la libertad de expresión. Sin embargo, dentro de la *deep web* existe un subconjunto más pequeño y oscuro conocido como *dark web*. La web oscura es una red oculta de sitios web que operan en redes superpuestas, como Tor (The Onion Router), que anonimizan las identidades de los usuarios y dificultan el rastreo del origen de la información. Aunque la web oscura es famosa por facilitar actividades ilegales como el tráfico de drogas, los servicios de piratería informática y los mercados ilegales, es importante señalar que no todas las actividades de la web oscura son ilícitas. Hay casos de uso legítimo, como la denuncia de irregularidades y la comunicación anónima, que dependen de la protección de quien la proporciona.

Vale la pena subrayar que acceder a la *deep web* o la *dark web* conlleva riesgos inherentes, como la exposición a actividades ilegales, imágenes perturbadoras, malware, estafas y pornografía.

Sobre el autor

Fabio Morandín-Ahuerma es Doctor (*cum laude*) en Filosofía por el Instituto de Filosofía de la Universidad Veracruzana y tiene una Maestría por el Instituto "Matías Romero" de la Secretaría de Relaciones Exteriores, así como Licenciatura en Filosofía por la Universidad Veracruzana. Realizó una Estancia Postdoctoral en el Centro de Investigaciones Filosóficas de Buenos Aires, Argentina (CIF) en su Programa de Neuroética Fundamental. Es profesor-investigador de tiempo completo definitivo de la Benemérita Universidad Autónoma de Puebla y líder del Cuerpo Académico BUAP-CA-354. Cuenta con publicaciones científicas y de divulgación en revistas nacionales y extranjeras. Ha sido profesor invitado en universidades en México y en el extranjero. Sus más recientes libros son: "Neuroética fundamental y teoría de las decisiones" (2021) y "Neuroeducación como herramienta epistemológica" (2022), ambos editados por el Consejo de Ciencia y Tecnología del Estado de Puebla (Concytep). Actualmente es Investigador Nacional Nivel 1 del Sistema Nacional de Investigadores del Consejo de Humanidades, Ciencia y Tecnología de México. Cuenta con Perfil Deseable PRODEP. Correo electrónico: fabio.morandin@correo.buap.mx. Página personal: https://fabiomorandin.net/

“Principios Normativos para una Ética de la Inteligencia Artificial”
Editado por el Consejo de Ciencia y Tecnología del Estado de Puebla

Fecha de publicación: septiembre 2023
Número de Páginas: 212

«PRINCIPIOS NORMATIVOS PARA UNA ÉTICA DE LA INTELIGENCIA ARTIFICIAL» ES UNA GUÍA EN ESPAÑOL SOBRE LAS CONSIDERACIONES ÉTICAS EN TORNO AL DESARROLLO, DESPLIEGUE Y USO DE LAS TECNOLOGÍAS DE IA EN LA ACTUALIDAD.

A TRAVÉS DE DIEZ CAPÍTULOS, SE RETOMAN LAS DIRECTRICES DE ORGANIZACIONES MULTINACIONALES COMO LA UNESCO Y LA OCDE, ASÍ COMO DE GRUPOS CENTRADOS EN LA IA COMO PARTNERSHIP ON AI (PAI) E IEEE, ENTRE OTROS. ESTE VOLUMEN ES EL RESULTADO DE UNA AMPLIA INVESTIGACIÓN QUE RESCATA LOS MARCOS DEONTOLÓGICOS DE LA IA, PARA DAR A LOS LECTORES UNA COMPRENSIÓN MATIZADA DE LOS PRINCIPIOS CLAVE QUE DEBEN GUIAR EL DESARROLLO ÉTICO Y EL USO DE LA TECNOLOGÍA DEL APRENDIZAJE AUTOMÁTICO.

ADEMÁS, EL AUTOR ANALIZA LAS NORMATIVAS DE ALGUNAS EMPRESAS DEL SECTOR QUE UTILIZAN Y DESARROLLAN IA COMO LO SON GOOGLE, MICROSOFT, FACEBOOK Y APPLE, OFRECIENDO A LOS LECTORES UNA VISIÓN PONDERADA DE SUS PROPUESTAS. LO MISMO SOBRE LOS MARCOS NORMATIVOS DE PAÍSES LÍDERES, ENFRASCADOS EN LA NUEVA “GUERRA FRÍA DE LA IA”: ESTADOS UNIDOS, CHINA Y RUSIA.

SI USTED DESARROLLA SISTEMAS DE IA, SI ES RESPONSABLE DE POLÍTICAS PÚBLICAS PARA REGULAR LA TECNOLOGÍA O ES UN ESTUDIANTE O CIUDADANO PREOCUPADO POR LAS IMPLICACIONES ÉTICAS DE LA IA, ESTE LIBRO ES UN RECURSO ESENCIAL DE ACCESO ABIERTO QUE OFRECE IDEAS Y ORIENTACIONES PERTINENTES.

EDITADO POR EL CONSEJO DE CIENCIA Y TECNOLOGÍA DEL ESTADO DE PUEBLA, ESTE TRABAJO SE UNE A OTROS TÍTULOS DEL MISMO AUTOR EN ESTA COLECCIÓN, COMO «NEUROÉTICA FUNDAMENTAL Y TEORÍA DE LAS DECISIONES» (2021) Y «NEUROEDUCACIÓN COMO HERRAMIENTA EPISTEMOLÓGICA» (2022).

PUEBLA Un gobierno presente
Secretaría de Educación
CONCYTEP Consejo de Ciencia y Tecnología del Estado de Puebla
Complejo Regional Nororiental