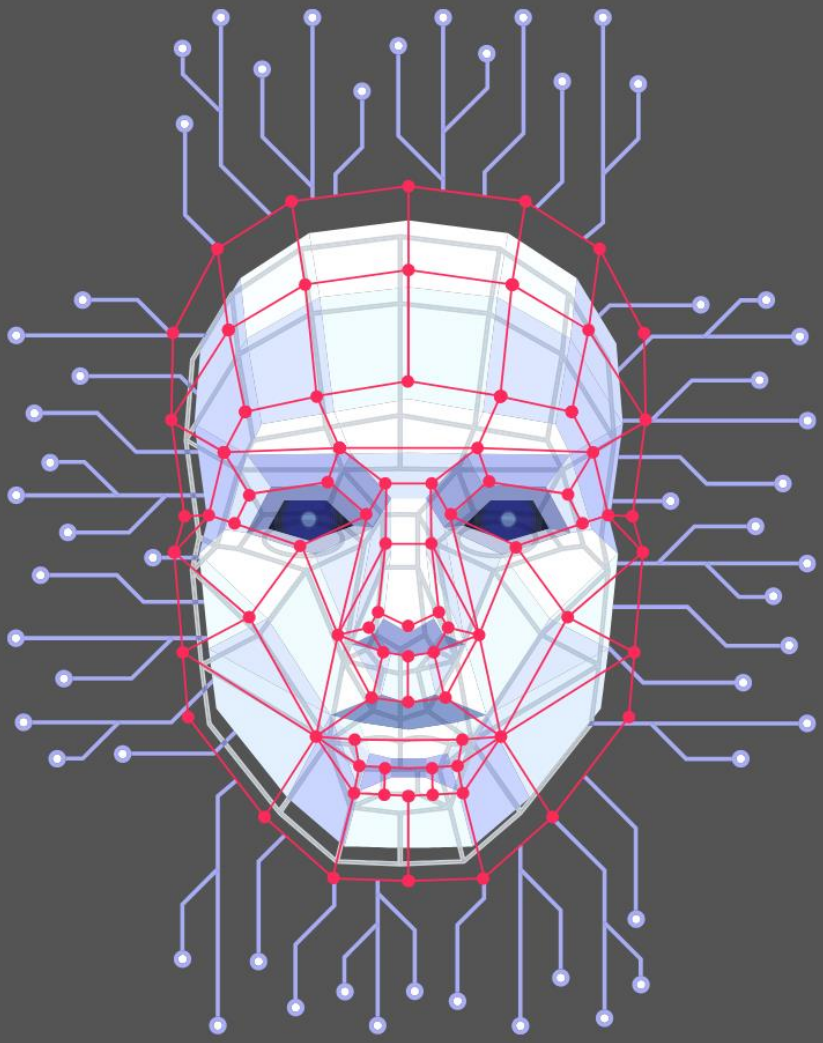
CONSEJO DE CIENCIA Y TECNOLOGÍA DEL ESTADO DE PUEBLA



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

ISBN: 978-607-8901-78-4
Primera edición, México, 2023

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.