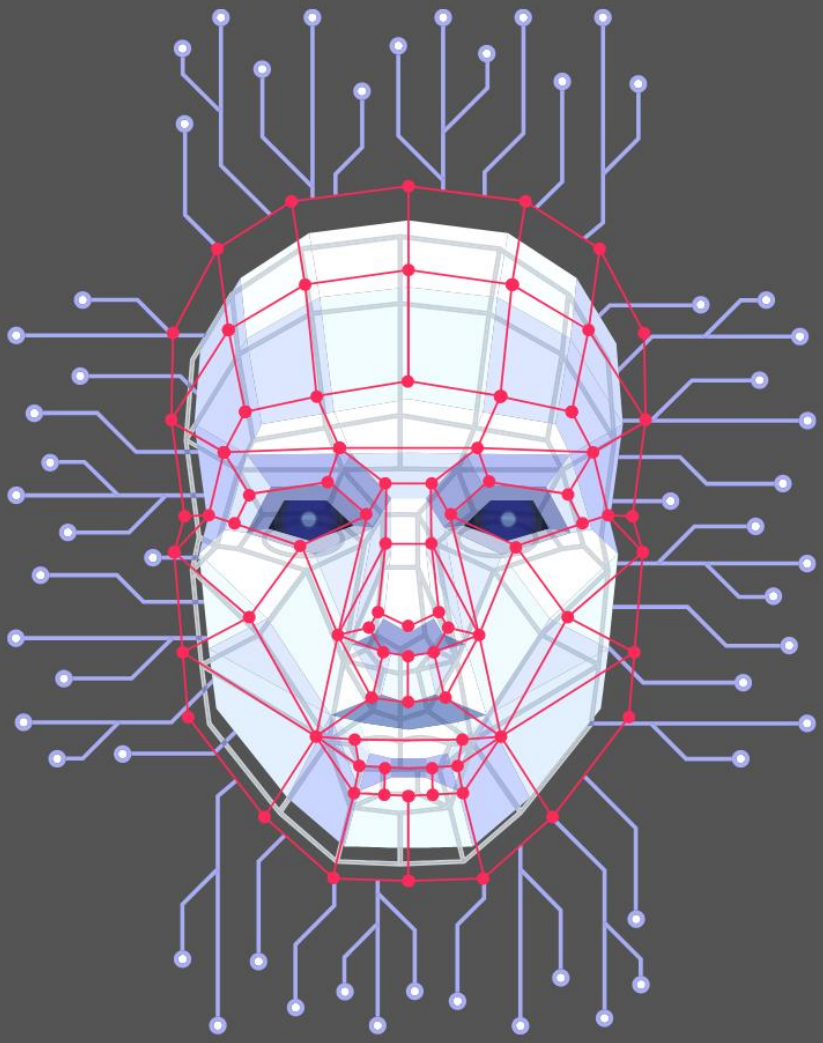
CONSEJO DE CIENCIA Y TECNOLOGÍA DEL ESTADO DE PUEBLA



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

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.